DOCUMENT RESUME

ED 209 353 TM 810 915

AUTHOR Reckase, Mark D.
TITLE The Formation of Homogeneous Item Sets When Guessing is a Factor in Item Responses.
INSTITUTION Missouri Univ., Columbia. Dept. of Educational Psychology.
SPONS AGENCY Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
REPORT NO ONR-RR-81-5
PUB DATE Aug 81
CONTRACT N00014-77-C-0097
NOTE 94p.

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Cluster Analysis; Difficulty Level; *Factor Analysis; *Guessing (Tests); *Latent Trait Theory; Mathematical Models; Multidimensional Scaling; Response Style (Tests); Simulation; *Test Construction; *Test Items

ABSTRACT

One of the major assumptions of latent trait theory is that the items in a test measure a single dimension. This report describes an investigation of procedures for forming a set of items that meet this assumption. Factor analysis, nonmetric multidimensional scaling, cluster analysis and latent trait analysis were applied to simulated and real test data to determine which technique could best form a unidimensional set of items. Theoretical and empirical evaluations were also made of the effects of guessing on the dimensionality of test data. The results indicated that guessing affected highly discriminating items more so than poorly discriminating items. Of the procedures evaluated for sorting items into unidimensional item sets, principal factor analysis of phi coefficients gave the best results overall. In summary, guessing does have an effect on test data, but the effect is not very large unless items of extreme difficulty are present in the test. Of the procedures evaluated, traditional factor analytic techniques gave the most useful information for sorting test items into homogeneous sets. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED209353

The Formation of Homogeneous Item Sets When Guessing is a Factor in Item Responses

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Mark D. Reckase

Research Report 81-5
August 1981

Tailored Testing Research Laboratory
Educational Psychology Department
University of Missouri
Columbia, MO 65211

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
The Office of Naval Research
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Prepared under contract No. N00014-77-C-0097,NR150-395
with the Personnel and Training Research Programs
Psychological Sciences Division
Office of Naval Research

Approved for public release; distribution unlimited.
Reproduction in whole or in part is permitted for any
purpose of the United States Government.

TM 815 935

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 81-5
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): The Formation of Homogeneous Item Sets When Guessing is a Factor in Item Responses
5. TYPE OF REPORT & PERIOD COVERED: Technical Report
6. PERFORMING ORG. REPORT NUMBER:

7. AUTHOR(s): Mark D. Reckase
8. CONTRACT OR GRANT NUMBER(s): N00014-77-6-0097

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
Department of Educational Psychology
University of Missouri
Columbia, MO 65211

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:
P.E.: 61153N Proj.: RR042-04
T.A.: 042-04-01
W.U.: NR150-395

11. CONTROLLING OFFICE NAME AND ADDRESS:
Personnel and Training Research Programs
Office of Naval Research
Arlington, Virginia 22217

12. REPORT DATE: August 1981
13. NUMBER OF PAGES: 80

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):

15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approval for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Latent Trait Theory
Dimensionality
Guessing
Similarity Coefficients
Factor Analysis
Multidimensional Scaling
Cluster Analysis

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

One of the major assumptions of latent trait theory is that the items in a test measure a single dimension. This report describes an investigation of procedures for forming a set of items that meet this assumption. Factor analysis, nonmetric multidimensional scaling, cluster analysis and latent trait analysis were applied to simulated and real test data to determine which technique could best form a unidimensional set of items.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Theoretical and empirical evaluations were also made of the effects of guessing on the dimensionality of test data. The results indicated that guessing affected highly discriminating items more so than poorly discriminating items. However, the effect of guessing on the dimensionality of tests with common distributions of difficulty and discrimination indices was found to be minimal. Of the procedures evaluated for sorting items into unidimensional item sets, principal factor analysis of phi coefficients gave the best results overall. Nonmetric multidimensional scaling also showed promise when used with Yule's Y, phi, or tetrachoric similarity coefficients, but it did not perform as well as the factor analytic techniques on the real test data. In summary, guessing does have an effect on test data, but the effect is not very large unless items of extreme difficulty are present in the test. Of the procedures evaluated, traditional factor analytic techniques gave the most useful information for sorting test items into homogeneous sets.

The Formation of Homogeneous Item Sets

When Guessing is a Factor in Item Responses

One of the fundamental assumptions of most latent trait models is that the items in the pool of interest measure a single latent trait (Lord and Novick, 1968). Although some item pools do approximate the conditions specified by this assumption (e.g., vocabulary, arithmetic computation, digit span, etc.), in many cases item pools do not automatically fulfill the requirements of a one-dimensional latent space. For example, most achievement tests designed using a table of specifications are not unidimensional. Further, it is questionable whether some criterion-referenced test item domains measure a single dimension. Therefore, some procedure is needed to form unidimensional item sets for use with latent trait models.

Unfortunately, the procedures commonly used to form item sets that are homogeneous in the ability measured have been criticized because of some basic inadequacies. Most of these criticisms stem from the use of items that are dichotomously scored. Factor analysis, for example, was derived for use with continuous variables. Since its basic model reproduces the observed score from a linear combination of continuous variables, there is no way that dichotomous responses can be adequately modeled. A symptom of this problem is the difficulty factors obtained when phi coefficients are factor analyzed. In an attempt to alleviate this problem, tetrachoric correlations are often used in place of phi coefficients. However, these correlations may not yield correlation matrices that have the appropriate properties for factor analysis (i.e., positive semidefinite). The end result of these problems is that the most commonly used multivariate sorting procedure is theoretically inadequate for forming unidimensional item sets when dichotomously scored items are used.

In response to the problems in the use of factor analysis with dichotomous variables, Christofferson (1975) has developed a factor analysis procedure specifically for this special case. In order to avoid the problems stemming from the use of correlation coefficients, he uses the proportions in two-way tables of item responses as the basic data for determining the factor structure. A generalized least squares procedure is used to estimate error free proportions, and from these, estimate the parameters of the factor analysis model. The obtained parameter estimates have been shown to be consistent and a chi-square test has been developed to test the number of significant factors. Although this procedure would seem to be the solution to the item factoring problem, it can only be used on a maximum of 25 items because of computer storage and computational time constraints. Thus, the procedure is not practical for most item pool construction situations.

Another approach has been taken by Divgi (1980) to solve the item pool dimensionality problem, but this procedure only provides a test for the presence of a single factor, rather than a procedure for sorting items. In Divgi's procedure, the probability of a correct response to an item (expected response) determined from a latent trait model is subtracted from the actual response to that item to obtain a residual. These residuals are then intercorrelated over items and the resulting correlation matrix is factor analyzed using

the principal components procedure. If any strong factors are left in the correlation matrix, it is proposed that this is evidence that unidimensionality does not hold. This procedure is purported to be better than the usual factor analysis of dichotomous variables because the correlations are based on the continuous residuals, rather than binary data. However, this procedure is very new and has not been critically evaluated. In any case, it does not yield a procedure for forming unidimensional item sets.

In addition to the above procedures for determining the dimensionality of item pools, cluster analysis and multidimensional scaling procedures are also available. These procedures make fewer assumptions, but their usefulness is unknown. Moreover, a review of the literature has not found any application of these procedures to the unidimensionality issue.

The end result of the confusion caused by the lack of good procedures for forming unidimensional item sets is that often item pools are sorted subjectively, without the aid of an analytic procedure. In many cases the dimensionality of the item pool is not checked at all. Obviously, an easily used procedure is needed to develop unidimensional item sets. One of the purposes of this research is to find such a procedure.

Unfortunately, the mere fact that dichotomously scored items are being used is not the only problem that affects the determination of the dimensionality of an item pool. For multiple-choice items, guessing is another factor that may affect the observed dimensionality. A review of the literature on latent trait theory and multivariate clustering procedures has found no studies on the effect of guessing on dimensionality, so the magnitude of these effects is unknown. However, work has been done on the effects of guessing on item analysis, correlation, and reliability. Some hints concerning guessing effects can be discovered there.

Carroll (1945) studied the effect of varied item difficulty and guessing on the magnitude of correlations between dichotomously scored items using the "knowledge or random guessing model". He found that variations in both difficulty and chance success bring about a reduction in the size of the phi coefficient between items. He also discussed the use of tetrachoric correlations with dichotomously scored test items, and showed that variations in difficulty had no effect on the tetrachoric correlations when no guessing was present and when the bivariate normal assumption was met. When guessing was present in the data the magnitude of the obtained correlations was lowered. This effect was stronger for more difficult items. Along with his analysis of the effects of guessing and difficulty on these two types of correlations, Carroll also developed correction formulae to compensate for the reduction in correlation. The correction for the tetrachoric correlation will be described later in this report, since it was used in the research reported here.

Plumlee (1952) expanded on Carrol's work to determine the effect of variation in difficulty and guessing on item-test correlations and reliability. She developed an equation in her article that showed the relationship between biserial correlations determined with and without guessing present, and another equation that showed the corresponding relationship for parallel form reliability. In both cases, the equations predicted a reduction in the magnitude of the statistics with the presence of guessing.

Plumlee then checked the accuracy of her equations by determining the item discrimination values and reliability using items administered in completion and multiple-choice form. The equations were used to predict the values for the statistics for the multiple-choice tests from the completion test statistics. The predictions were close, but there was a tendency to over estimate the statistics. The differences were explained by the inaccuracy of the "knowledge or random guessing model".

Mattson (1965) also determined the effects of guessing on reliability, but he used a different approach than Plumlee. Mattson used a binomial error model to estimate the standard error of measurement and the true score variance. He then showed how the true score variance is reduced by guessing effects. From the standard error and true score variance terms, he developed a formula for the reliability of a test when guessing is a factor. The reliability was shown to decline with increased guessing probability.

A totally different approach to the determination of the effects of guessing on reliability was taken by Denney and Remmers (1940). They felt that the addition of choices to a multiple-choice item was, in fact, analagous to lengthening the test. Thus, the reliability of the test with more alternative choices could be determined from the test with fewer choices using the Spearman-Brown formula (a four choice test is twice as long as a two choice test). In their article they present data that showed that the Spearman-Brown formula does model the guessing effect fairly well. In that study, vocabulary items were administered with two, three, four, or five choices and the reliability was determined for each of the test forms using the split-half method. In their article, as in all of the others, the reliability decreased with increased guessing.

To summarize the various theoretical positions, the proportion of true variance in a set of test scores was plotted against guessing level for a test with a no-guessing reliability of .81. The results are shown in Figure 1. Four plots are shown on this graph. The first is the predicted reliability of a test as a function of guessing for a test with a no-guessing reliability of .81. This plot was produced using Equation 30 developed by Carroll (1945). A 50 item test composed of items with .50 traditional difficulty was assumed in making this plot. The second plot shows the effects of guessing on the squared biserial correlation between an item of .5 traditional difficulty and total score when the no-guessing correlation is .9. This relationship was determined from Equation 24 in the article by Plumlee (1952). The third line shows the relationship between reliability and guessing given in Table 1 from an article by Mattson (1965). A no-guessing reliability of .8 was assumed for this plot. The fourth line shows the reliability as a function of guessing level as determined by the method proposed by Denney & Remmers (1940). The values were derived using the generalized Spearman-Brown formula, assuming a reliability of .81 for a test composed of items with 10 alternatives.

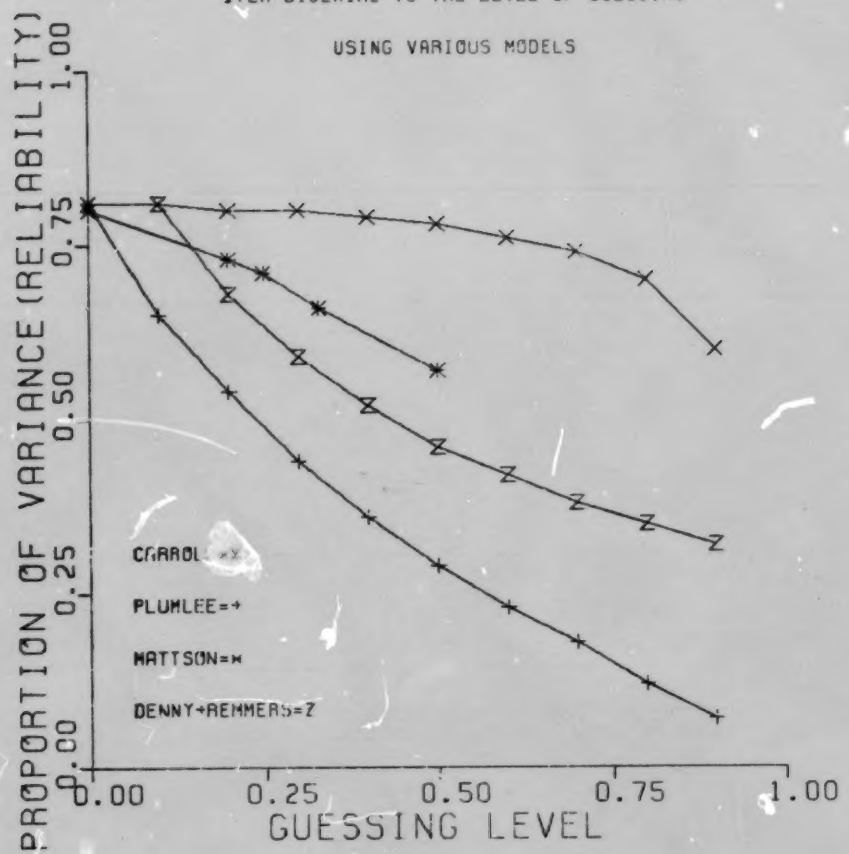
As can be seen from this figure, the predicted reliabilities are quite different. Other than concluding that the reliability declines, no consistent prediction can be made about the magnitude of the decline. The implication of these data to the proportion of common variance in a test is that guessing effects will

FIGURE 1

RELATIONSHIP OF TEST RELIABILITY AND

ITEM BISERIAL TO THE LEVEL OF GUESSING

USING VARIOUS MODELS



cause the common variance to decrease. The lower correlations suggested by Carroll's work would also imply that the number of factors (in a factor analytic sense) would increase.

Since no clear cut findings were discovered in the review of the literature concerning the effects of guessing on multidimensional data reduction techniques, the present research study was designed to further explore these effects. More specifically, the purpose of the research was to evaluate various procedures for forming homogeneous item sets, and to determine the effects of guessing on the techniques. Three approaches were taken to achieve this goal. First, a theoretical model was developed, and guessing effects were predicted with the model. Secondly, simulated data were generated using the theoretical model, and the predicted results were checked by actual analysis of these data. Third, a real data-set was selected and analyzed to determine how well the theoretical and simulated results generalized. Conclusions were drawn from consistent patterns of findings from these three sets of results.

The Theoretical Model

The basic model used here to determine the effects of guessing on the proportion of common variance in an item is a modification of the true score model presented in Lord and Novick (1968, pp. 30-38). A univariate model will be presented first, followed by a multivariate generalization.

Suppose that a population of examinees is normally distributed on a unidimensional trait, T, that is required, to some extent, for performance on a test item. Without loss of generality, this distribution can be assumed to have a mean of zero and a variance of one. That is, $T \underline{d} N(0, 1)$. Suppose further that the trait measured by a test item, T', is not exactly the same as trait T, but that it has a positive relationship with T. If the correlation between the person trait, T, and the trait measured by the item, T', is given by "a", then the score on trait T' for Person j, t_j, can be estimated from his/her score on T, τ_j, by the formula

$$t_j = a\,\tau_j \qquad (1)$$

if a linear relationship is assumed, and if T' is assumed to have a standard normal distribution---that is, $T' \underline{d} N(0, 1)$.

If the item in question yields responses on a continuous scale, the observed score on the item is given by the usual true score model as

$$x_j = t_j + \varepsilon, \qquad (2)$$

where x_j is the observed score on the item for Person j, t_j is the person's true score on the trait defined by the item score, and ε is a random error term which is distributed $\varepsilon \underline{d} N(0, \sigma^2), \sigma > 0$, for Person j.

Based on Equations 1 and 2,

$$E(x_j) = E(t_j) + E(\varepsilon) \qquad (3)$$
$$= E(a\tau_j) + 0$$
$$= a\tau_j,$$

since trait estimate τ_j is constant for Person j. Since $E(x_j)$ is the classical definition of a true score, the true score on the item is defined as $t_j = a\tau_j$.

The variance of the observed score on Person j on the item is given by

$$V(x_j) = V(t_j + \varepsilon)$$
$$= V(t_j) + V(\varepsilon) + 2\mathrm{cov}(t_j, \varepsilon).$$

Since t_j is constant for Person j, and since the covariance with error is assumed to be zero,

$$V(x_j) = 0 + V(\varepsilon) = \sigma^2. \qquad (4)$$

Up to this point the expectation and variance of the observed score, x_j, has been obtained based on the probability distribution of scores for a single person. Similar results can also be determined for the entire population of individuals. Notationally this will be indicated by starring the subscript indicating the person. The expectation of the score on the item is then given by

$$E(X_\star) = E(T_\star') + E(\varepsilon)$$
$$= E(aT_\star) + 0$$
$$= a\,E(T_\star)$$
$$= a \cdot 0 = 0. \qquad (5)$$

The variance of scores on the test item is given by

$$V(X_\star) = V(T_\star') + V(\varepsilon) + 2\mathrm{cov}(T_\star', \varepsilon)$$
$$= V(aT_\star) + \sigma^2 + 0$$
$$= a^2 \cdot V(T_\star) + \sigma^2$$
$$= a^2 + \sigma^2, \qquad (6)$$

since $T_\star$ has a variance of 1.0 and the covariance of the trait score and error is assumed to be zero. If the item trait scores are assumed to be

in standard score form, $V(X_*) = a^2 + \sigma^2 = 1.0$. Therefore,

$$\sigma^2 = 1 - a^2.$$

Equation 4 can then be written as

$$V(x_j) = 1 - a^2. \tag{7}$$

Since the real interest of this report is the effects of guessing on the factor structure of dichotomously scored tests, the continuous item score will now be dichotomized by specifying a value c related to the difficulty of the item. If x_j is greater than c, a score of 1.0 will be assigned to Person j, and if x_j is less than c, a score of 0.0 will be assigned. More concisely,

$$\text{if} \quad x_j > c, \; u_j = 1$$

$$\text{and if} \quad x_j \leq c, \; u_j = 0$$

where u_j is the dichotomous score for the item for Person j.

The probability that a person with ability τ_j will get a score of $u_j = 1$ on the item is

$$P(U_j = 0|\tau_j) = \int_{z_c}^{+\infty} \phi(z)dz, \tag{8}$$

where $z_c = \dfrac{c - E(x_j)}{\sqrt{V(x_j)}} = \dfrac{c - a\tau_j}{\sqrt{1 - a^2}}$, and $\phi(z)$ is the normal probability density function. The probability of a score of $u_j = 0$ for a person with ability τ_j is

$$P(U_j = 0|\tau_j) = \int_{-\infty}^{z_c} \phi(z)\, dz.$$

This is essentially the normal ogive IRT model.

If Person j obtains a score of 0 on the item, (i.e., he/she does not know the correct answer), he/she may guess the correct answer with probability 1/A, where A is the number of alternatives in the item if it is assumed to be multiple-choice. That is, with a 1/A probability, the 0 will be changed to a 1. Therefore, the probability of obtaining a score of 1 on the item when guessing is a factor is given by

$$P'(U_j = 1|\tau_j) = P(U_j = 1|\tau_j) + P(U_j = 0|\tau_j) \cdot 1/A. \tag{9}$$

One way to conceptualize the effect of guessing on this item is that guessing causes the cutting score, c, to be shifted downward, increasing the probability of a correct response. To determine the magnitude of this shift, the cutting socre, c', that yields the correct probability of a correct response, including the guessing effect, can be determined using the inverse normal transformation:

$$z'_j = \phi^{-1}(1 - P'(U_j = 1|\tau_j)). \tag{10}$$

The value of c' is obtained by transforming this z-score to the observed score scale using

$$c'_j = z'_j \sqrt{1 - a^2} + a\tau_j. \tag{11}$$

Note that c' has an index, j, denoting that its value may be different for each person, depending on the ability level τ_j. The guessing effect for Person j can then be defined as

$$g_j = c - c_j. \tag{12}$$

Another way of conceptualizing the effect of guessing is that it shifts upward the examinee's propensity distribution by an amount g_j.

Based on the idea of guessing causing a shift in a person's propensity distribution, a new continuous score for an item can be defined to include guessing as a factor in the item response:

$$y_j = x_j + g_j = t_j + \varepsilon + g_j, \tag{13}$$

where g_j is constant for Person j on any given item, but varies across people and items. The expected value of this score for Person j is

$$\begin{aligned} E(y_j) &= E(t_j) + E(\varepsilon) + E(g_j) \\ &= E(a\tau_j) + 0 + g_j \\ &= a\tau_j + g_j. \end{aligned} \tag{14}$$

In classical test theory, this expected value is defined as the true score on the item for Person j (Lord & Novick, 1968). Notice that there is a guessing component in this classical true score. The variance of y for Person j is given by

$$\begin{aligned} V(y_j) &= V(t_j) + V(\varepsilon) + V(g_j) + 2cov(t_j, g_j) + 2cov(\varepsilon, g_j) + 2cov(t_j, \varepsilon) \\ &= 0 + 1 - a^2 + 0 + 0 + 0 + 0 \\ &= 1 - a^2 \end{aligned} \tag{15}$$

since t_j and g_j are constant and the covariance with error is assumed to be zero.

The probability of a correct response to an item when ability is measured on the y-scale (i.e., when guessing is a factor in the item response) is given by

$$P'(U_j = 1|\tau_j) = \int_{z'_j}^{+\infty} \phi(z)dz, \tag{16}$$

where

$$z'_j \quad \frac{c - a\tau_j + g_j}{\sqrt{1 - a^2}}.$$

As was done previously, these results can be generalized to apply to the scores obtained from a group of individuals rather than for a single individual, as given in Equations 8 through 16. The expected value of the continuous item score for the population of individuals when guessing is a factor in responses is

$$\begin{aligned} E(Y_\star) &= E(T'_\star) + E(\varepsilon) + E(G_\star) \\ &= E(aT_\star) + 0 + E(G_\star) \\ &= a\,E(T_\star) + E(G_\star) \\ &= 0 + E(G_\star) = E(G_\star), \end{aligned} \tag{17}$$

where $G_\star$ is the random variable associated with the guessing effect. Thus, the average score on the item for the population is increased over the no-guessing score by an amount equal to $E(G_\star)$. The variance of the Y-score for the population is given by

$$\begin{aligned} V(Y_\star) &= V(T'_\star) + V(\varepsilon) + V(G_\star) + 2\mathrm{cov}(T'_\star, G_\star) + 2\mathrm{cov}(\varepsilon, G_\star) \\ &= V(aT_\star) + 1 = a^2 + V(G_\star) + 2\mathrm{cov}(aT_\star, G_\star) + 0 + 0 \\ &= a^2V(T_\star) + 1 - a^2 + V(G_\star) + 2a\mathrm{cov}(T_\star, G_\star) \\ &= 1 + V(G_\star) + 2a\mathrm{cov}(T_\star, G_\star). \end{aligned} \tag{18}$$

From Equations 17 and 18, the proportion in the population that will obtain a score of $U_\star = 1$ can be determined. This proportion is given by

$$P'(U_* = 1) = \int_{z'_c}^{\infty} \phi(z)dz, \tag{19}$$

where

$$z'_c = \frac{c_i - E(G_*)}{1 + V(G_*) + 2acov(T_*, G_*)}.$$

The development of this model has now reached the point where it can be applied to the major area of interest of this paper---determining the effect of guessing on the proportion of variance accounted for by the common factors in a test. First, for the unifactor case, the proportion of variance accounted for on the item in question by the unidimensional ability, τ, when no guessing is present can be obtained from Equation 6 and the expression for the variance of the true scores, t, over the population of interest:

$$V(T'_*) = V(aT_*) = a^2\ V(T_*) = a^2. \tag{20}$$

The proportion of observed variance for the item accounted for by the true scores is then

$$\frac{V(T'_*)}{V(X_*)} = \frac{a^2}{a^2 + \sigma^2} = \frac{a^2}{a^2 + 1 - a^2} = a^2. \tag{21}$$

Thus, the proportion of variance accounted for by the item trait is simply the squared correlation between the trait and the true score on the item.

The proportion of variance accounted for by the item true scores when guessing is a factor in item responses is given by the ratio

$$V(E(y_j))/V(Y_*).$$

The numerator of this ratio, the variance of the true scores, can be obtained from Equation 14 as

$$\begin{aligned} V(E(y_j)) &= V(aT_j + G_j) \\ &= V(aT_j) + V(G_j) + 2cov(aT_j, G_j) \\ &= a^2\ V(T_j) + V(G_j) + 2acov(T_j, G_j) \\ &= a^2 + V(G_*) + 2acov(T_j, G_j). \end{aligned} \tag{22}$$

Using the value for variance of the observed score given by Equation 18, the ratio of the true score variance to the observed score variance is given by

$$\frac{V(E(y_j))}{V(Y_*)} = \frac{a^2 + V(G_*) + 2a\mathrm{cov}(T_*, G_*)}{1 + V(G_*) + 2a\mathrm{cov}(T_*, G_*)} . \qquad (23)$$

In the univariate case, Equations 21 and 23 simply give the reliability of a single test item.

The results for the univariate case can be generalized to the multivariate case by redefining t_j as

$$t_j = \sum_{k=1}^{m} a_k \tau_{jk}, \qquad (24)$$

where $\tau_{jk} \underline{d} N(0, 1)$ for each j and k, the a_k are the correlations between the τ_{jk} and the continuous score on the test items, and m is the number of abilities required to perform on the items. The τ_{jk} and τ_{jl} are assumed to be uncorrelated for $k \neq l$. The proportion of common variance in the no guessing case then becomes

$$\frac{V(T'_*)}{V(X_*)} = \frac{V(\Sigma a_k T_{*k})}{1} = \frac{\sum_{k=1}^{m} a_k^2 V(T_{*k}) + \sum_{i \neq j}^{m} a_i a_j \mathrm{cov}(T_{*i}, T_{*j})}{1}$$

$$= \frac{\sum_{k=1}^{m} a_k^2}{1} = h^2, \qquad (25)$$

where h^2 is the communality. When guessing is a factor, Equation 23 becomes

$$\frac{V(E(Y_j))}{V(Y_*)} = \frac{\sum_{k=1}^{n} a_k^2 + V(G_*) + 2\sum_{k=1}^{n} a_k \mathrm{cov}(T_{*k}, G_*)}{1 + V(G_*) + 2\sum_{k=1}^{n} a_k \mathrm{cov}(T_{*k}, G_*)} . \qquad (26)$$

Predictions from the Theoretical Model

With the development of the theoretical model presented on the previous pages, it is possible to determine the magnitude of guessing effects for persons of a given ability, and for populations with known distributions of ability, assuming the "knowledge or random guessing model" is correct. For example, if an individual with known ability $\tau_j = -1$ is administered an item with difficulty .5 for the population as a whole, a guessing level of .05, and a correlation between the item and trait T of .9, several important features can be determined. First, the expected score on the trait defined by item performance can be obtained from Equation 1 as -.9. The variance of the estimate on the item trait for the person is given by Equation 7 as .19.

Based on a cut score of 0.0 for the population for the no-guessing case, the probability that Person j obtains a correct response to the item can be obtained from Equation 8 as .019. After introducing the effect of guessing into this item, this person's probability of a correct response is .069 (from Equation 9). This change in probability requires a shift in the person's propensity distribution of 1.49 standard deviation units, yielding a guessing effect from Equations 11 and 12 of .252.

This same procedure can be followed for all levels of ability in the population. If the probability distribution of ability in the population is known, an expected guessing level for the population as a whole can be determined using

$$E(G) = \int_0^\infty g\, f(g)\, dg \qquad (27)$$

Unfortunately, g has a functional form that contains the inverse normal function, so direct computation of the expected value is impossible. Therefore, for the purposes of this report, E(G) has been computed using the cautious adaptive Romberg extrapolation method (IMSL, 1979) of numerical integration.

Table 1 gives the magnitude of the expected value of the guessing effect for combinations of the probability of guessing on the item and the correlation between the item trait, t, and the person trait, τ. The probability of guessing is defined here as the probability of a correct response for a person with no knowledge of the material measured by the item. The correlation between the item and person traits is the same as the loading of the item on the first factor of a test measuring the person trait. A cutting score of 0.0 was used for all combinations of guessing level and factor loading because any other cutting score would yield a simple linear transformation of these results.

Most of the results presented in Table 1 match what would commonly be expected of a guessing effect. As the probability of guessing increased, the guessing effect increased. However, for low guessing probabilities the guessing effect increased with increased factor loading, while for high guessing probabilities, the guessing effect decreased with increased factor loading. At a guessing probability of approximately .25, the guessing effect was fairly constant. This interaction of guessing probability and factor loading was unanticipated.

Table 1

Expected Value of the Guessing Effect for $c_i = 0$, and Various Combinations of Guessing Level and Correlation Between Item Trait and Person Trait

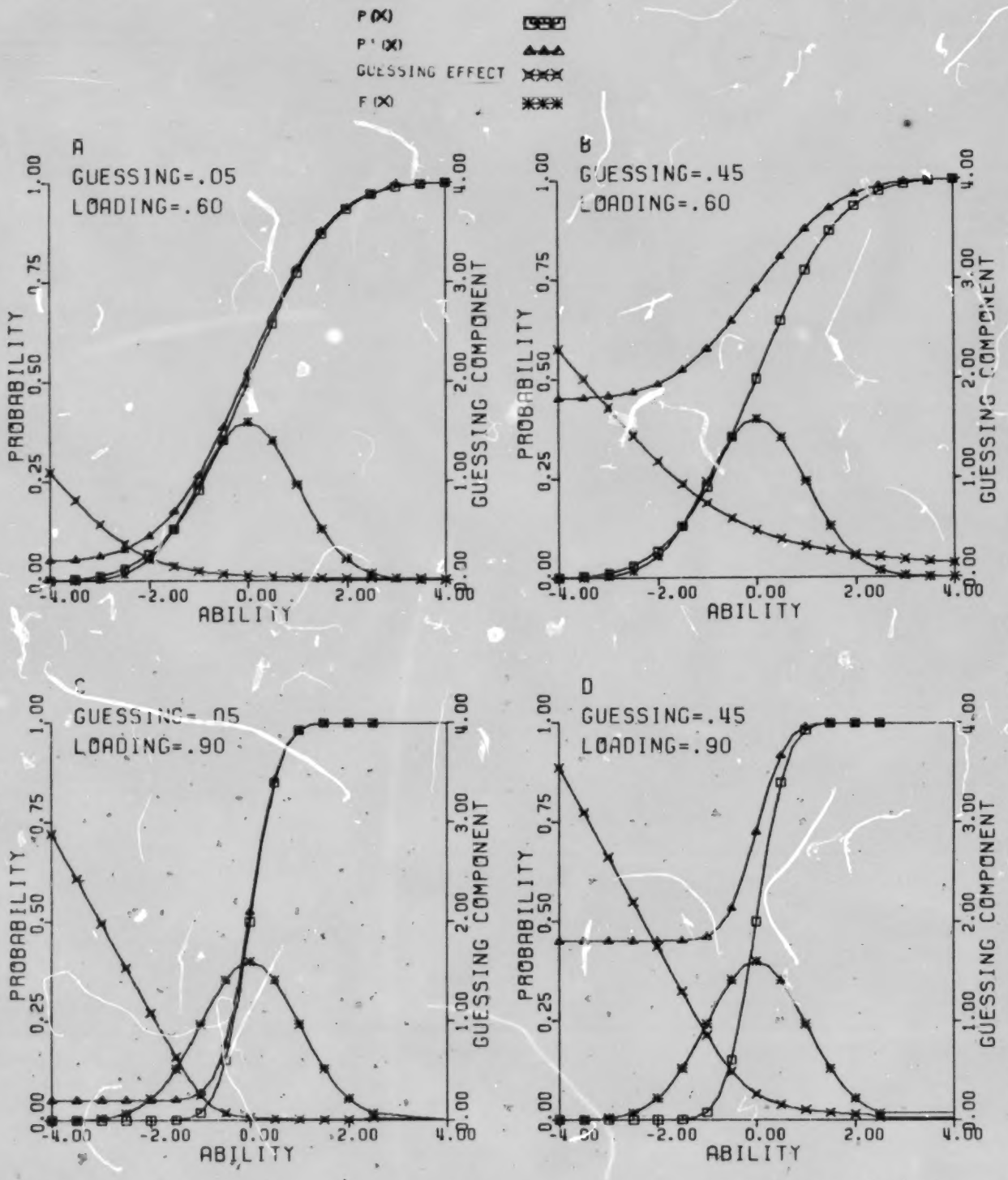
Guessing Level	Correlation Between Person Trait and Item Trait			
	.6	.7	.8	.9
.05	.07	.08	.10	.14
.15	.19	.19	.21	.24
.25	.30	.30	.30	.31
.35	.41	.40	.39	.38
.45	.53	.51	.48	.45
.55	.66	.62	.58	.51
.65	.80	.75	.68	.59
.75	.98	.91	.81	.68

Note. Expected values were based on a N(0, 1) ability distribution.

The reason for this interaction can be determined from Figures 2a, 2b, 2c, and 2d, which show the probability of a correct response to the item with and without guessing, the guessing effect at various ability levels, and the ability density function for guessing probabilities and first factor loading of (.05, .6), (.05, .9), (.45, .6) and (.45, .9), respectively. From these figures it can be seen that the magnitude of guessing increases more quickly with decrease in ability for the .9 loading case than the .6 loading case. This yields a higher expectation for the .9 loading case with a .05 guessing level than for the .6 loading case, because the guessing effect reaches an appreciable size within the range containing most of the ability distribution in the former case, but not in the latter. When the guessing probability is .45, the higher guessing level over the entire ability range for the .6 loading item overcomes the steeper slope of the guessing effect for the .9 loading item. In other words, the guessing effect is greater for the poorer item over a wider range of ability.

FIGURE 2

PLOT OF THE PROBABILITY OF CORRECT RESPONSE WITH AND WITHOUT GUESSING EFFECT AND ABILITY DISTRIBUTIONS FOR VARIOUS GUESSING LEVELS AND FACTOR LOADING



From a practical point of view, these results suggest that guessing at reasonable levels is a more serious problem for high quality items than low quality items. In the latter case, the error variance in the item masks the guessing effects. Of course, this conclusion assumes the correctness of the "knowledge or random guessing model".

Although the magnitude of the guessing effect has resulted in some interesting findings, the more important area of interest in this report is the reliability, or proportion of common variance as a function of guessing level. This value can be determined from Equation 23, but first the variance of the guessing effect, and the covariance of the guessing effect and ability are required. The formulae used to obtain these statistics using numerical integration are given by:

$$V(G) = \int_0^\infty (g - E(g))^2 \, f(g)dg,$$

and

$$\text{cov}(\tau, G) = \int_{-\infty}^{\infty} \tau(g - E(g)) \, f(\tau)d\tau.$$

The expression in the equation for the covariance is integrated over τ, since g is a function of τ. It should be recalled that $\tau \underline{d} N(0, 1)$.

Table 2 gives the variance and covariance values for the same probability of guessing values and the level of factor loadings used in Table 1. From this table it can be seen that as the guessing level increases, the variance increases, and the covariance of guessing and ability becomes more negative. Also, the same trend can be seen as the factor loading increases.

The negative covariances were expected in these results, since low ability individuals guess more often the high ability individuals. The increase in variance was also expected. As the guessing level increases, the guessing effect function shown in Figures 2a through 2d is shifted upward, demonstrating a greater range of guessing effect. With increased factor loading, the guessing effect function increases more sharply, resulting in the greater magnitude of the variance. It is not surprising that as the variance increases the covariance also increases in absolute value.

From the variance of the guessing effect, the covariance of guessing and ability, as well as the factor loading, the proportion of variance in item responses accounted for by ability can be determined from Equation 23. These proportions are presented for the cases used in Tables 1 and 2 in Table 3. In addition to the 0.0 cutting score case (corresponding to .5 traditional difficulty for the group), the proportions are presented for the .75 and .25 traditional difficulty cases.

Table 2
Variance of the Guessing Effect and
the Covariance of the Guessing Effect and the
Trait Level for C_i = k and
Various Combinations of Guessing Level
and Factor Loadings

Guessing Level		First Factor Loading			
		.6	.7	.8	.9
.05	VAR	.00	.01	.03	.08
	COV	-.05	-.07	-.12	-.20
.15	VAR	.02	.03	.06	.12
	COV	-.11	-.15	-.20	-.28
.25	VAR	.03	.05	.09	.15
	COV	-.15	-.20	-.25	-.33
.35	VAR	.04	.07	.11	.18
	COV	-.19	-.24	-.30	-.37
.45	VAR	.06	.09	.13	.21
	COV	-.23	-.28	-.33	-.40
.55	VAR	.07	.11	.16	.23
	COV	-.26	-.31	-.36	-.43
.65	VAR	.09	.13	.18	.25
	COV	-.29	-.34	-.40	-.46
.75	VAR	.11	.15	.21	.28
	COV	-.32	-.37	-.43	-.49

Table 3

Proportion of Variance Accounted for in Item Responses by Guessing Level, Factor Loading, and Cutting Score

Cutting Score	Factor Loading	Guessing Level								
		.00	.05	.15	.25	.35	.45	.55	.65	.75
0	.9	.81	.73	.69	.66	.64	.61	.59	.56	.53
	.8	.64	.57	.51	.47	.44	.40	.37	.34	.31
	.7	.49	.44	.38	.34	.31	.27	.24	.22	.19
	.6	.36	.32	.28	.24	.21	.18	.16	.14	.12
-.6745	.9	.81	.80	.78	.77	.76	.74	.73	.72	.70
	.8	.64	.62	.59	.57	.55	.53	.50	.48	.45
	.7	.49	.47	.45	.42	.40	.38	.35	.33	.30
	.6	.36	.35	.32	.30	.28	.26	.24	.22	.20
.6745	.9	.81	.60	.52	.47	.43	.39	.36	.33	.29
	.8	.64	.45	.36	.30	.26	.22	.19	.17	.14
	.7	.49	.35	.26	.21	.17	.14	.12	.10	.08
	.6	.36	.26	.18	.14	.11	.09	.07	.06	.05

Note. Ability is assumed to be distributed N(0, 1).

Note first that as the guessing level increases, the proportion of item variance accounted for by ability decreases, and that it decreases more dramatically for the more difficult items. For the most difficult item (cutting score of .6745), even the .05 guessing level has a substantial effect. As the size of the factor loading declined, the proportion of variance in the item scores accounted for by ability also declined, as expected.

It must be kept in mind when interpreting these results that they refer to the proportion of variance accounted for by ability (reliability) for only one item. The values for a test made up of many items would be substantially higher, the actual value depending on the distribution of difficulties of the items, their individual guessing levels, and the interitem covariances. Because of the complexity of the problem of determining the reliability of a test using the theoretical model proposed, only the reliability for a single item is presented.

Evaluation of Empirical Item Sorting Procedures

Since a theoretical analysis of the effects of guessing on the proportion of common variance in a test composed of many items was not possible, the more realistic, and therefore more complex, cases were studied by applying the available item sorting procedures to various simulated data-sets and real data-sets. The basic design for this component of the research study was to produce item sets with known structure using simulated and real test results, and then attempt to recover the structure using each of several available techniques. The techniques considered included: factor analysis, cluster analysis, nonmetric multidimensional scaling, and latent trait theory.

Besides the choice of techniques to be used on the item response data, another decision needed to be made concerning the coefficient used as a measure of similarity between the items. Factor analysis is rather limited in this choice, being tied to correlation type statistics. Cluster analysis and nonmetric multidimensional scaling do not have this limitation, opening up the possibility of using many other measures of similarity. Therefore, the following coefficients were applied to the data, and the various techniques were applied to each: phi coefficient, tetrachoric correlation, corrected tetrachorics, eta coefficient, Yule's Q, Yule's Y, approval score, Kendall's tau B, Goodman/Kruskal's gamma, agreement score, and the Lijphart index. The formula and reference for each of these coefficients is given in Appendix A.

Item Sorting Procedures

Each of the techniques used in the analysis of the item response data has many variations in basic procedure as well as several options as to specific method of application. Therefore, before describing the research design any further, the specific techniques used will be described to make their identity unambiguous.

Factor Analysis Two basic factor analysis procedures were used on the data: the method of principal components, and the method of principal factors. These two procedures differ mainly in that the former assumes that all of the variance influences the magnitude of the correlations, while the latter assumes that some variance is unique to each item and a reduced number of factors (less than the number of items) explains the correlations. Although the latter procedure seems more reasonable, both were used on selected data-sets to determine their relative value.

In addition to the basic factor analysis procedures, two types of rotations were used to help in the interpretation of the results. The two rotations used were VARIMAX, and OBLIMIN. These two were selected because of their general availability, and because they allowed comparisons between orthogonal and oblique solutions.

The factor analyses were run on only three of the similarity coefficients mentioned above: the phi coefficient, the tetrachoric correlation, and the tetrachoric correlation corrected for guessing (Carroll, 1945). The other coefficients were not used because they did not even approximate the assumptions of the factor analytic model.

Because of the different factor analytic options available in different packages, four different packages were used to perform the analyses. These included SPSS (Nie, Hull, Jenkins, Steinbrenner and Brent, 1975), SOUPAC (Computing Services Office, 1974), OSIRIS III (Institute for Social Research, 1974) and SAS (Barr, Goodnight, Sall, and Helwig, 1976). In some cases, the same analyses were run using two different packages to check their comparability. Differences in results obtained from the different package programs were minor.

Cluster Analysis Two different cluster analysis approaches were taken for this study. The first, labeled CLUSTER for this report, builds clusters of items one at a time. The procedure first searches the input similarity matrix for the two items with the highest similarity. The matrix is then searched for the item that has the highest minimum similarity to the three items in the cluster. This item is also added to the cluster. This procedure continues until no items can be found with a similarity greater than a pre-set cut-off value. At that point the matrix is again searched for the two items not included in the first cluster with the highest similarity. These two items form the beginning of a new cluster. The clustering procedure then continues until all items are used or until none of the similarities exceed the cut-off value.

The second clustering procedure used, called HICLUSTER in this report, is a hierarchical clustering procedure. In this procedure, the most similar pair of items is connected first, then the next most similar, and so on, to form initial clusters. These initial clusters are combined when all of the points in one cluster are connected to all of the points in another cluster. Clustering in this procedure continues until all of the items are included in one cluster. All of the similarity coefficients listed above were used with both of these procedures.

Both of the clustering procedures used for this study were applied using programs from the OSIRIS III computer program package. Although this package is not as widely available as SAS or SPSS, the clustering routines from this package were used because of their greater versatility.

Nonmetric Multidimensional Scaling The nonmetric multidimensional scaling procedure used for the data analysis in this study was the basic MDSCAL procedure developed by Shepard (1962) and Kruskal (1964). This procedure rank orders the similarity of the items used in terms of the specified similarity coefficient used, and then attempts to define a space of minimum dimensionality such that the distances between the items in the space are ranked in the same order as the initial similarities. The procedure uses a steepest descent iterative approach to improve the relationship between the spatial configuration and the initial similarities. When the rate of improvement levels off, the solution is accepted. A Euclidean metric was used for all of the analyses using this procedure.

The OSIRIS III version of MDSCAL was used for all of the analyses presented in this report. Each of the coefficients mentioned above was used as a similarity measure for this procedure, since it makes no assumption other than an ordinal scale concerning the coefficients. Although numerous other multidimensional scaling algorithms exist, this algorithm was selected because it is widely available.

Latent Trait Analysis Although latent trait analysis is not commonly thought of as a multidimensional clustering technique, some results obtained in previous research suggested that it might be used as such (Reckase, 1979). That research suggested that when several factors are present the LOGIST (Wood, Wingersky, & Lord, 1976) item calibration program selects one factor as a basis for item calibration. Thus, items with high discrimination parameter estimates should be from the same latent dimension, while those with low estimates should be from other dimensions. By deleting the highly discriminating items after each run of the program, another set of highly discriminating items may be found that measure a different latent dimension. Thus, iterative use of the program, with item deletions between successive iterations, may yield sets of homogeneous items. It was the purpose of the analyses performed for this research to determine if that were indeed the case.

Data-Sets

As mentioned earlier, the item sorting procedures were applied to two kinds of data-sets: simulated and actual test data. The simulated responses of examinees to items were used so that precise control could be maintained over the dimensionality of the data. The actual test data were used to get a more realistic evaluation of the procedures. The production procedures and characteristics of the data-sets will now be described.

Simulated data A total of 24 simulated data-sets were produced for this study. These data-sets all represented the responses to 50 items by 1000 individuals. They varied in the number of dimensions used to generate the responses, the distribution of item difficulties, the guessing level, and the distribution of the guessing level. All of the data-sets were generated using a variation of the procedure described by Wherry, Naylor, Wherry & Fallis (1965).

This procedure generates data using the basic linear factor analytic model. A more detailed description of the procedure is given in Reckase (1979).

The procedure developed by Wherry et. al. generates data to match a specified factor structure, but does not include a guessing effect. Therefore, after the simulated responses were produced using the above procedure, the incorrect responses to an item were randomly changed to correct responses at a rate equal to the guessing probability for the item. This was done by comparing a flat random number on the 0.0 to 1.0 range with the guessing level for the item, and changing an incorrect response to a correct response if the selected random number were less than the guessing level.

The total list of data-sets produced for the study are presented in Table 4. As can be seen from the table, more than half of the data-sets produced used only one generating factor. These data-sets were produced to determine the effect of guessing on the obtained dimensionality of a set of test data. Both the level of guessing and the distribution of parameters were varied for these data-sets.

The next set of data-sets listed in the table used two orthogonal factors to generate the item responses. This set of relatively simple multidimensional data-sets was used to determine which procedure could adequately find the homogeneous item sets within the test. If a procedure were not successful on this "easy" set of data, it was eliminated from consideration.

The remaining simulated data-sets used three or nine orthogonal factors to generate the item responses. These data-sets were generated to have a large first factor to more accurately simulate what was believed to be a realistic state of nature. Only item sorting procedures that succeeded on the two-dimensional data were applied to these more complex data-sets.

Real Data The real data-set used in this study was produced by sampling items and responses from the results of the 1975-76 administration of the Iowa Tests of Educational Development (1972). The desire here was to produce a test with two underlying dimensions that contained all the sources of variation present in typical test administrations. To achieve the desired dimensionality, items were selected from the Expression and Quantitative Thinking subtests of the ITED. These two subtests were judged to be most dissimilar, and so most likely to yield the desired structure. A total of 50 items were randomly selected from the 105 items in the two subtests using a stratified random sampling approach. Thirty-three of the items in this data-set were from the Expression subtest and 17 were from the Quantitative Thinking subtest.

Table 4

Catalogue of Data-Sets

Dimensionality of Data Set	Labels*
1-Factor	SD15ON.CG00, SD15OR.CG00, SD15OR.CG05
	SD15OR.CG15, SD15OR.NG15, SD15ON.CG15
	SD15OR.CG25, SD15OR.NG25, SD15ON.CG25
	SD15ON.CG35, SD15OR.CG35, SD15ON.CG45
	SD15OR.CG45, SD15ON.CG55, SD15OR.CG55
	SD15ON.CG65, SD15OR.CG65, SD15ON.CG75
	SD15OR.CG75
2-Factor	SD25OR.CG00, SD25ON.NG20, SD25OR.CG25
3-Factor	SD35ON.NG20
9-Factor	SD95ON.NG20

* The label of the data-set describes the data-set. The first two letters stand for simulation data. The next three or four digits tell the number of factors and number of items. All data-sets contained 50 items. The letter following the 50 tells the distribution of traditional item difficulties: N or R meaning normal or rectangular, respectively. Following the period is CG or NG, standing for constant or normally distributed guessing. The final two digits give the guessing level. The values given are the guessing level for CG data-sets or the mean guessing level for NG data-sets.

Research Design

The basic research design of this study contains four components. First, the four techniques of interest (factor analysis, cluster analysis, nonmetric multidimensional scaling, and latent trait analysis) were applied to the one dimensional data-sets, with guessing varied, to attempt to discover the effect of guessing on the techniques. This was done by plotting various characteristics of the techniques (e.g., size of first eigenvalue) against the guessing level to determine if any relationship existed. Also, the structure of the data-sets was considered unknown and the results of the procedures were analyzed to determine if the unidimensional structure could be discovered. This set of analyses formed a basis for comparison for all of the subsequent analyses.

The second analysis component consisted of applying the four techniques to the two, three, and nine-dimensional data-sets. For each of the data-sets, an attempt was made to recover the underlying structure of the data. If a procedure failed for a low dimensional data-set, it was not used with the more complex data-sets.

The third analysis component consisted of applying the four techniques to the real data-set. The procedure used with the real data-set was similar to that used with the simulated data-sets. The techniques were evaluated on their ability to reproduce what was thought to be the underlying structure of the data-set. In this case the data-set was constructed to have two components, but since the true structure could not be determined with certainty the interpretation of the results was much more cautious.

The final analysis performed for this study was the comparison of the results obtained using the simulation data with those suggested by the research literature. That is, the obtained reliability as a function of guessing was compared to the theoretical predictions.

Results

One-Dimensional Simulated Data

The results of the application of the four techniques to the one-dimensional simulated data will be presented first. The factor analysis results will be presented first, followed by the multidimensional scaling, cluster analysis, and latent trait analysis results.

Factor Analysis The first analysis performed using the factor analysis procedure was determination of the relationship between the size of the first factor on the test and the magnitude of the guessing component contained in the responses to each item on the test. To obtain this information, a principal components factor analysis was performed on tetrachoric correlations for eight data-sets. These data-sets were all generated using a normal distribution of traditional item difficulty centered around .5. Each was generated using a

constant guessing level. The guessing level used were 0, .15, .25, .35, .45, .55, .65, and .75. All data-sets were generated so that each item had a .9 loading on the first factor before the guessing effect was added.

To show the relationship between the guessing level and the size of the first factor, the proportion of total test variance accounted for by the first factor was plotted against the size of the guessing component. This plot is given in Figure 3, along with a plot of the KR-20 reliability against the guessing level. As can be seen from the plot, the proportion of variance accounted for by the first factor dropped off substantially with an increase in guessing. At the .15 guessing level, the proportion of variance had already declined to .62 from the .83 obtained for the no guessing case. It is interesting to note that the decline in the KR-20 reliability is not nearly as dramatic, showing its insensitivity to guessing effects.

Along with the analysis of the proportion of variance accounted for by the first factor, an attempt was also made to determine if guessing induced additional factors in the test. That is, did the decline in the first factor indicate the presence of other factors. To determine this, the number of factors in each factor analysis was determined using the skree technique. The factor loadings for those factors were than studied to determine whether they were interpretable. For all cases except the .00 and .75 guessing level data-sets, two factors seemed to be present in the data. The second factor for all of the two factor cases looked like a guessing factor, with high loadings for the difficult items. For higher guessing levels, the second factor was not as clear, disappearing altogether for the .75 guessing level data-set.

In addition to the creation of a second factor in these data-sets, the loadings of the items on the first factor were also affected. They were found to decline with an increase in the difficulty of the test items. Since this did not occur for the .00 guessing data-set, the effect can be attributed to guessing.

Since the guessing factors were defined by loadings on the hard items, it would seem reasonable that the distribution of item difficulties would interact with the guessing effect. To test this conjecture, a data-set was produced with a rectangular distribution of difficulty rather than a normal one, as in the previous data-sets, and a .25 guessing level. The results of the principal component analysis of this new data-set showed that the presence of items of more extreme difficulty had the effect of reducing the proportion of variance accounted for by the first factor from .50 to .41 and increasing the number of factors in the data-set. The shree technique indicated four factors, but only three were readily interpreted. The second and third factors for this data-set both seemed to be guessing factors. For comparison purposes, the factor loadings for the first two principal components from the normally distributed data-set, and the first three from the rectangularly distributed data-set are presented in Table 5. Notice the guessing factors in the data and the decline in the first factor loadings with the increased difficulty of the items. Several other data-sets were produced with rectangularly distributed difficulties and their analysis produced similar results.

FIGURE 3

PROPORTION OF VARIANCE

IN THE FIRST FACTOR

AND RELIABILITY

AS A FUNCTION OF GUESSING LEVEL

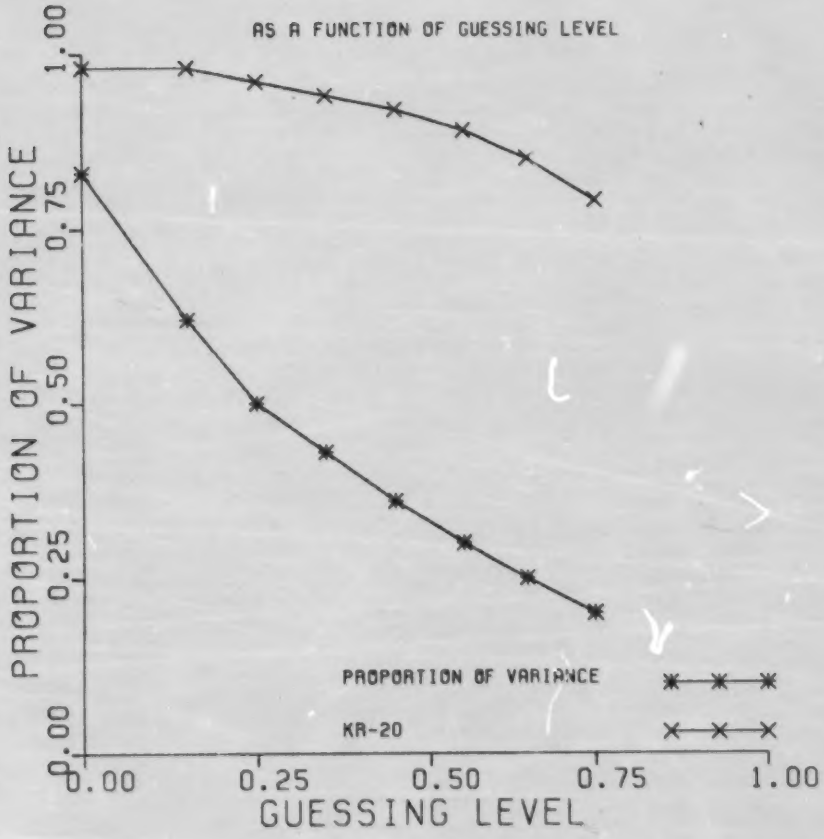


Table 5
Principal Component Factors from Tetrachoric Correlations
for Simulated Tests with Normal and Rectangular Distributions
of Difficulty and .25 Guessing Level

Item	Difficulty	Factors (Normal Distribution)		Difficulty	Factors (Rect. Distribution)		
		I	II		I	II	III
1	15	39	34	01	06	13	31
2	18	44	37	03	09	23	07
3	19	55	33	05	17	15	36
4	21	50	48	07	12	29	42
5	22	59	22	09	20	36	31
6	27	60	27	11	25	25	19
7	29	64	21	13	30	27	38
8	29	58	37	15	29	43	05
9	29	64	24	17	41	23	37
10	31	66	26	19	39	31	32
11	33	68	21	21	53	25	14
12	34	68	20	23	49	33	18
13	34	71	18	25	47	36	-14
14	37	65	18	27	54	32	14
15	38	68	17	29	55	40	-03
16	39	74	13	31	63	31	-01
17	41	68	16	33	55	36	-04
18	42	73	-01	35	66	24	-15
19	42	67	13	37	62	32	-05
20	47	75	12	39	66	28	-07
21	48	73	07	41	63	34	-26
22	48	73	01	43	68	26	01
23	52	72	03	45	66	28	-11
24	52	72	-06	47	74	19	-07
25	52	72	-01	49	71	20	-19

Note. All values are presented without decimal points.

Table 5 (Continued)
Principal Component Factors from Tetrachoric Correlations for Simulated Tests with Normal and Rectangular Distributions of Difficulty and .25 Guessing Level

Item	Factors (Normal Distribution) Difficulty	I	II	Factors (Rect. Distribution) Difficulty	I	II	III
26	54	77	02	51	72	18	-14
27	54	74	02	53	70	08	06
28	54	73	-03	55	76	10	-11
29	55	73	08	57	73	14	-13
30	55	75	-07	59	75	09	-14
31	56	78	-04	61	78	-10	17
32	57	78	-10	63	72	06	-21
33	58	75	-08	65	77	-09	-07
34	58	74	06	67	79	-08	-09
35	58	77	-07	69	75	-09	-15
36	60	76	-04	71	81	-18	-01
37	60	76	-18	73	77	-08	-07
38	60	74	-14	75	78	-14	-10
39	61	74	-10	77	78	-20	-13
40	61	79	-18	79	78	-25	-05
41	62	77	-15	81	77	-23	-01
42	64	77	-18	83	77	-23	-05
43	64	78	-21	85	79	-32	-01
44	65	80	-24	87	79	-35	13
45	65	75	-26	89	77	-39	01
46	66	74	-27	91	77	-36	-04
47	69	78	-32	93	76	-37	19
48	70	77	-35	95	79	-49	17
49	70	75	-35	97	75	-43	02
50	79	71	-42	99	63	-60	55

Note. All values are presented without decimal points.

From this set of analyses, three types of results were observed concerning the effects of guessing on the unidimensional simulation data. First, guessing reduced the contribution of the first principal component to the factor analysis results. Second, the loadings of the items on the first factor were reduced to the extent that guessing affected the items. Third, extra factors which seemed to be guessing factors were present in the principal component solution. Similar results were obtained when the principal factor procedure was used instead of the principal component solution.

Since the purpose of this report is to find methods for recovering unidimensional sets of items from a test, one further analysis was run on the one-factor rectangular distribution of difficulty data-set with .25 guessing. The purpose of this analysis was to determine if the correlation matrix could be corrected for guessing. Carroll's (1945) correction for the four-fold tables used to compute tetrachoric correlations was selected for this purpose. Since the true guessing level for an item is not usually known, the data were corrected for guessing using .15, .25, and .35 guessing levels. The corrected tetrachoric correlation matrices were then factor analyzed using the principal component technique. The first two factors obtained for the corrected matrices and the uncorrected solution are shown in Table 6.

The most obvious result that can be seen in Table 6 is that overcorrecting for guessing (.35 correction) results in a very unusual factor analysis solution. The first seven items defined unique factors, and many of the factor loadings were essentially 1.0. Overcorrecting for guessing clearly does serious harm to the factor analysis, resulting in meaningless results.

Correcting for guessing at the .15 and .25 level gave more reasonable results. The first factor loadings were increased above the uncorrected values. In many cases the .25 correction yielded loadings close to the .9 values used to generate the data. The .15 correction did little to remove the second factor from the solution. The .25 correction did tend to restrict the influence of the second factor to fewer items, mainly the most difficult items in the data-set.

In general, these results indicate that the correction for guessing for the tetrachoric correlations has some merit, but care must be taken not to overcorrect. The first factor loadings are improved by the procedure, but the correction did not totally remove the second factor, which was attributed to guessing.

Nonmetric multidimensional scaling The first analysis performed using the nonmetric multidimensional scaling technique was the application of the MDSCAL program to the one factor data with a rectangular distribution of item difficulty and no guessing. This analysis was performed on the data using each of thirteen similarity coefficients. These included the following: agreement coefficient, koppa coefficient, kappa coefficient, Lijphart index, Kendall's tau B, approval score, phi coefficient, Yule's Q, Yule's Y, phi over phi max, gamma, tetrachoric correlation, and eta coefficient. Each of these coefficients is described in Appendix A. This large set of coefficients was used since MDSCAL does not require any special characteristics in a measure of similarity, and it was hoped that one of these coefficients would be less sensitive to guessing effects than the others.

Table 6
Factor Pattern Matrices for a Two-Factor
Principal Component Solution of One-Factor Data with .25 Guessing
with Various Levels of Correction for Guessing

Item	No Correction		.15 Correction		.25 Correction		.35 Correction	
	I	II	I	II	I	II	I	II
1	06	13	14	24	64	75	00	00
2	09	23	17	42	78	92	00	00
3	17	15	36	18	99	48	00	00
4	12	29	24	51	66	63	00	00
5	20	36	32	59	71	69	00	00
6	25	25	38	36	75	30	00	00
7	30	27	49	34	91	14	00	00
8	29	43	41	56	64	51	97	-61
9	41	23	60	22	90	09	101	07
10	39	31	[illegible]	37	72	33	100	-03
11	53	25	75	15	100	-06	101	07
12	49	33	67	30	95	-06	101	07
13	47	36	60	34	79	-01	99	-48
14	54	32	71	28	95	12	101	07
15	55	40	70	40	94	10	101	07
16	63	31	79	26	100	00	101	07
17	55	36	66	35	80	17	99	-29
18	66	24	81	10	98	-14	101	08
19	62	32	74	26	87	19	100	03
20	66	28	78	23	94	13	101	07
21	63	34	73	28	82	13	98	-39
22	68	26	78	24	93	20	101	04
23	66	28	75	25	86	19	98	-04
24	74	19	86	09	97	03	101	-01
25	71	20	80	13	93	08	99	-23

Note. All values are presented without decimal points.

Table 6 (Continued)
Factor Pattern Matrices for a Two-Factor
Principal Component Solution of One-Factor Data with .25 Guessing
with Various Levels of Correction for Guessing

	No Correction		.15 Correction		.25 Correction		.35 Correction	
Item	I	II	I	II	I	II	I	II
26	72	18	82	06	90	-01	99	07
27	70	08	79	01	89	06	97	11
28	76	10	85	03	96	00	101	09
29	73	14	80	09	91	12	98	-18
30	75	09	82	03	91	02	97	-18
31	78	-10	85	-12	92	-08	99	06
32	72	06	78	00	84	-14	95	-29
33	77	-09	83	-18	88	-21	98	07
34	79	-08	84	-14	91	-17	99	11
35	75	-09	80	-20	82	-43	96	09
36	81	-18	87	-25	90	-48	99	08
37	77	-08	82	-07	92	14	96	-14
38	78	-14	82	-17	87	-22	97	05
39	78	-20	82	-24	88	-17	96	-11
40	78	-25	82	-31	84	-54	97	08
41	77	-23	81	-25	86	-13	96	08
42	77	-23	81	-25	89	-23	96	04
43	79	-32	83	-34	87	-46	98	07
44	79	-35	84	-31	93	-14	98	05
45	77	-39	80	-35	86	-24	97	15
46	77	-36	80	-37	88	-29	98	-04
47	76	-37	82	-28	94	-03	97	04
48	79	-49	83	-46	95	-43	99	06
49	75	-43	84	-41	85	-71	98	06
50	63	-60	81	-37	72	-52	80	1.16

Note. All values are presented without decimal points.

After the MDSCAL analysis was completed, the resulting two-dimensional configurations were plotted and the stress of the solutions were noted. Stress is a measure of the deviation of the obtained distances between the items in the MDSCAL solution from the distances present in the initial data. The value is standardized by the squared deviation of all of the distances from the mean distance. The smaller the stress, the better the fit of the MDSCAL solution.

The results of the analysis of the coefficients applied to the one-factor data indicated that three different types of solutions were being obtained for the data. Six of the coefficients (agreement, koppa, kappa, Lijphart, tau B, and phi) yielded plots that placed the items along a straight line on one dimension, with the items ordered in difficulty--the easy items at one end and the difficult items at the other. The values of the stress index varied from .048 to .029, with the kappa coefficient giving the smallest value. The reason for this pattern is that these coefficients are all affected by the difficulty of the test items, with items close together in difficulty being judged more similar. A plot of the MDSCAL result for the kappa coefficient is given in Figure 4.

The second type of solution was obtained for six of the coefficients (Yule's Q, Yule's Y, phi/phi max, gamma, tetrachoric, and eta). This solution resulted in a circular cluster of points. The position of the items within the cluster seemed to have no obvious relationship to the difficulty of the items. The stress value for the solutions ranged from .34 to .33, with Yule's Q and gamma giving the smallest values. This solution is a result of the fact that these coefficients are not affected by the difficulty of the item, and therefore all pairs of items are found to be equally similar for one-dimensional data. The circular pattern is a result of trying to get all of the items equal distances apart in a two-dimensional space. Of course this cannot be done, so the stress of the solutions are higher than those for the first set of coefficients. An example of the circular solution for the gamma coefficient is given in Figure 5.

The third type of solution obtained from the MDSCAL procedure resulted from the application of the approval statistic. This solution had all of the easy items clustered tightly in the center, with the hard items spread out to one side. The pattern is a result of the way this statistic is computed. It is simply the proportion of times both items are answered correctly at the same time. Thus, easy items are found to be more similar than hard items, which have fewer correct responses. This solution had the lowest stress of all of the procedures, with a value of .021. The plot of this solution is given in Figure 6.

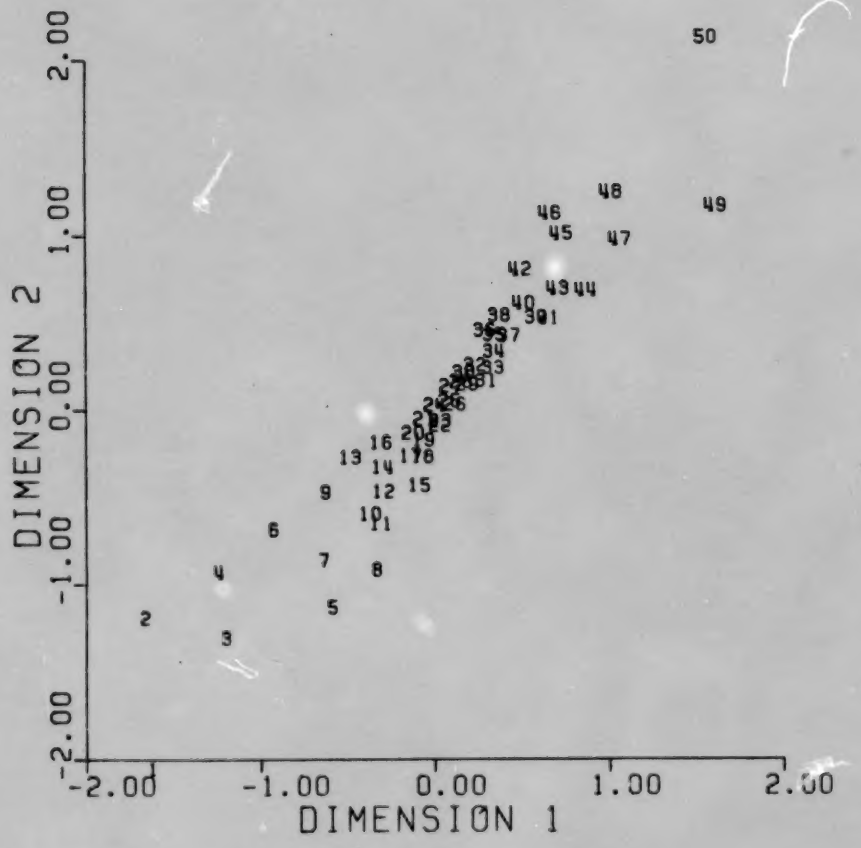
The next analysis run using the nonmetric multidimensional scaling technique was the computation of the two-dimensional solution using each coefficient for the one-factor, rectangular distribution of difficulty data-set with a .25 guessing level. The purpose of this analysis was to determine which coefficient gave a solution that was least affected by guessing.

FIGURE 4

TWO-DIMENSIONAL MDSCAL SOLUTION

FOR THE ONE-FACTOR NO GUESSING DATA

USING THE KAPPA COEFFICIENT



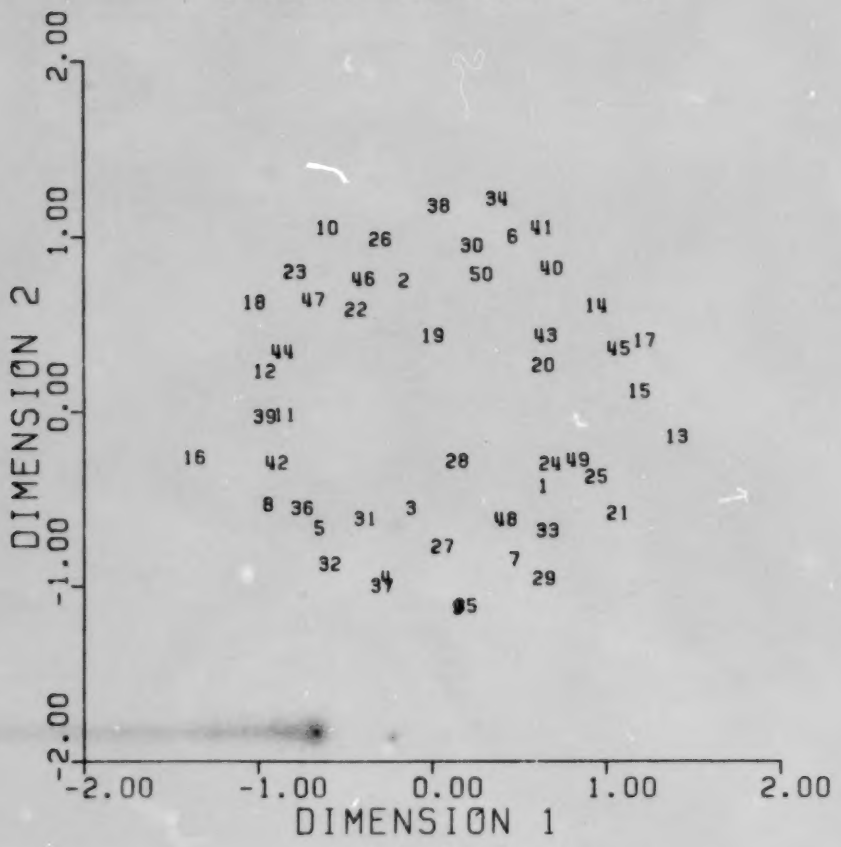
NOTE. THE ITEMS ARE NUMBERED WITH 1 THE MOST DIFFICULT AND 50 THE EASIEST

FIGURE 5

TWO-DIMENSIONAL MDSCAL SOLUTION

FOR THE ONE-FACTOR NO GUESSING DATA

USING THE GAMMA COEFFICIENT



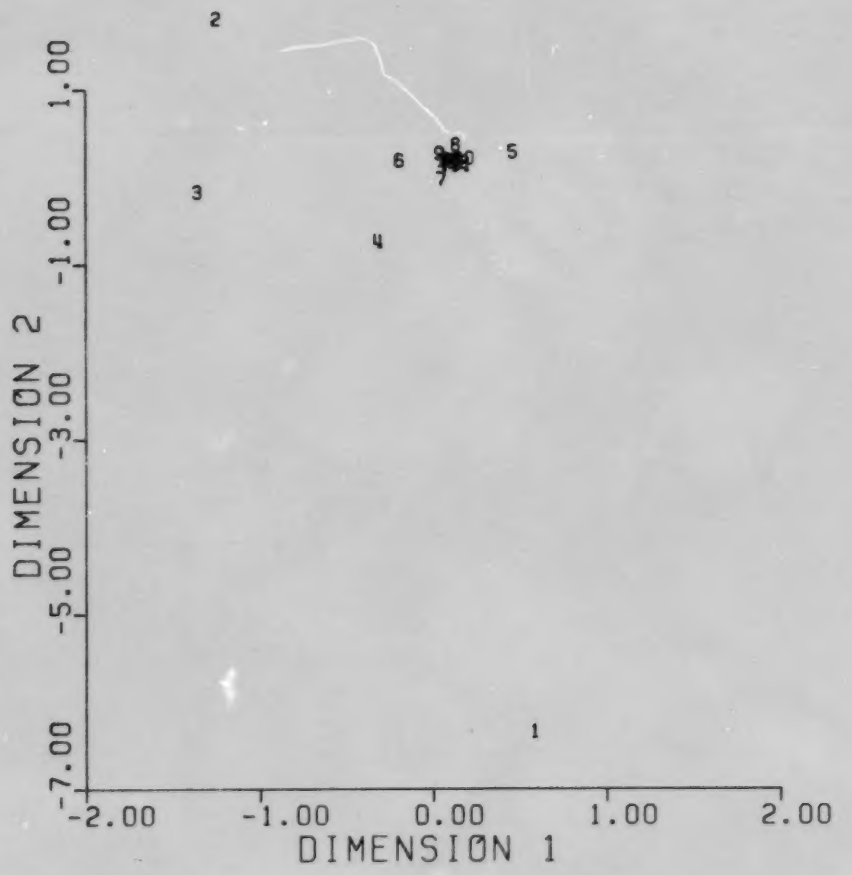
NOTE. THE ITEMS ARE NUMBERED WITH 1 THE MOST DIFFICULT AND 50 THE EASIEST

FIGURE 6

TWO DIMENSIONAL MDSCAL SOLUTION

FOR THE ONE-FACTOR NO GUESSING DATA

USING THE APPROVAL COEFFICIENT



NOTE. THE ITEMS ARE NUMBERED WITH 1 THE MOST DIFFICULT AND 50 THE EASIEST

The coefficients that yielded linear plots for the no guessing case gave two different types of plots for the .25 guessing case. The agreement, kappa and Lijphart coefficients resulted in wedge shaped plots for the two dimensional MDSCAL solutions, with the items high in guessing being in the wide part of the wedge. The stress was identical for all three coefficients, with a value of .104. This was substantially higher than the .042 achieved for the no guessing case. Figure 7 shows the plot of the results for the agreement score. Since these three coefficients gave identical results, only the agreement score will be given further consideration.

The second type of plot obtained from the coefficients giving a linear plot for the no guessing data was a crescent shape with the easiest and hardest items at the points of the crescent. The Kendall's tau B, phi and kappa coefficients gave this type of pattern. The stress for these solutions ranged from .137 to .168, up from .029 to .048 when no guessing was present. Of these three coefficients, kappa resulted in the MDSCAL solution with the smallest stress value. The plot of the two-dimensional solution for the kappa coefficient is given in Figure 8. The effect of guessing on these coefficients seems to be an increased similarity in the very easy and very hard items, resulting in the curvature in the plots.

The coefficients that resulted in circular patterns for the one-dimensional data with no guessing also yielded two patterns when MDSCAL was applied to the one-dimensional data with .25 guessing. The Yules Q, Yule's Y, phi/phi max, gamma, and tetrachoric coefficients all resulted in two-dimensional solutions that showed the original circular patterns distorted by pulling the items most affected by guessing down to the lower left. Guessing increased the distance between the items most affected by guessing, causing greater dispersion for those items. The stress values for the solutions ranged from .219 to .270, with the tetrachoric correlation giving the smallest value. Phi/phi max gave the largest stress value and showed the greatest dispersion for the easy items. The distortions caused by guessing brought about a reduction in the stress value from the .33 value obtained when no guessing was present. It seems that the guessing effect brings about a more linear continuum than was present previously, making the data easier to fit. The plot of the two-dimensional MDSCAL solution for the tetrachoric correlations is presented in Figure 9.

The second type of solution obtained from the set of coefficients that resulted in circular plots was obtained for the eta coefficient. In this case the plot remained circular, but the hard items migrated to the circumference of the two-dimensional structure, while the easy items moved to the center. The stress for this solution increased from .330 for the no guessing solution to .365 for the guessing solution. Figure 10 presents the plot of this solution.

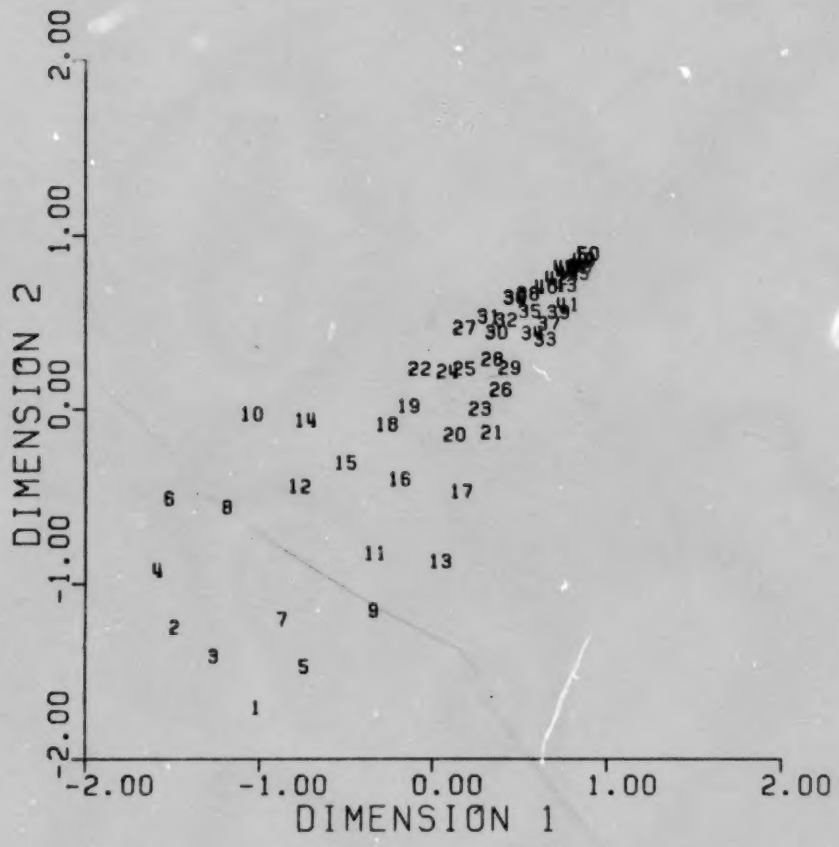
The approval score, the coefficient that gave the third type of pattern for the no guessing data, resulted in a pattern similar to that obtained for the eta coefficient when guessing was present. A circular pattern resulted, with the hard items at the circumference and the easy items at the center. The center cluster was much tighter in this case, however. The stress of this

FIGURE 7

TWO DIMENSIONAL MDSCAL SOLUTION

FOR THE ONE-FACTOR .25 GUESSING DATA

USING THE APPROVAL COEFFICIENT



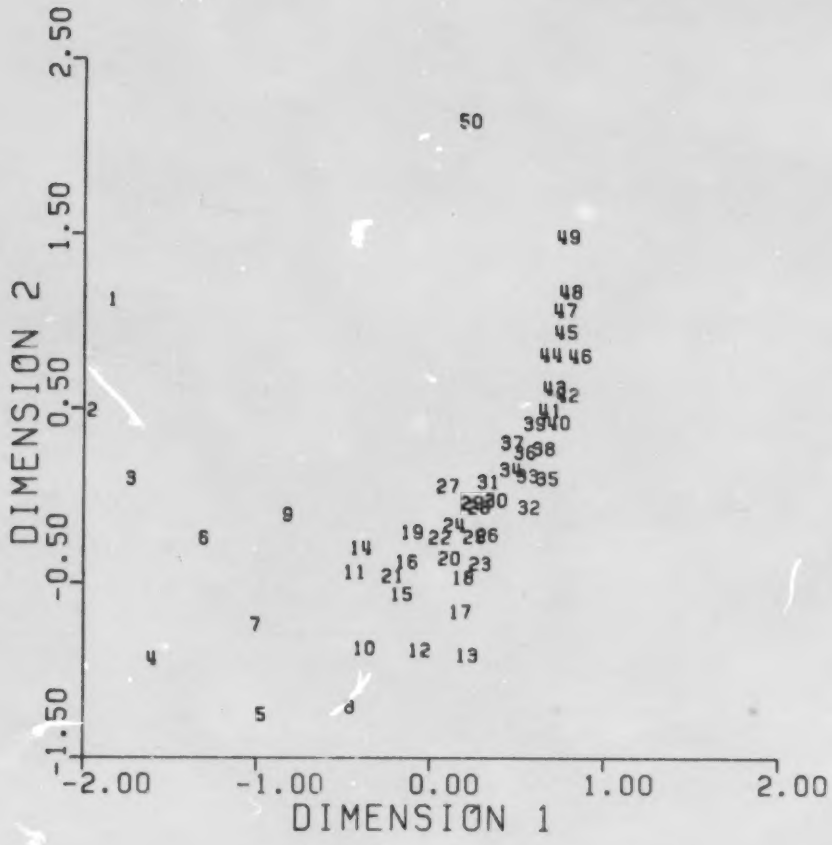
NOTE. THE ITEMS ARE NUMBERED WITH 1 THE MOST DIFFICULT AND SO THE EASIEST

FIGURE 8

TWO-DIMENSIONAL MDSCAL SOLUTION

FOR THE ONE-FACTOR .25 GUESSING DATA

USING THE KAPPA COEFFICIENT



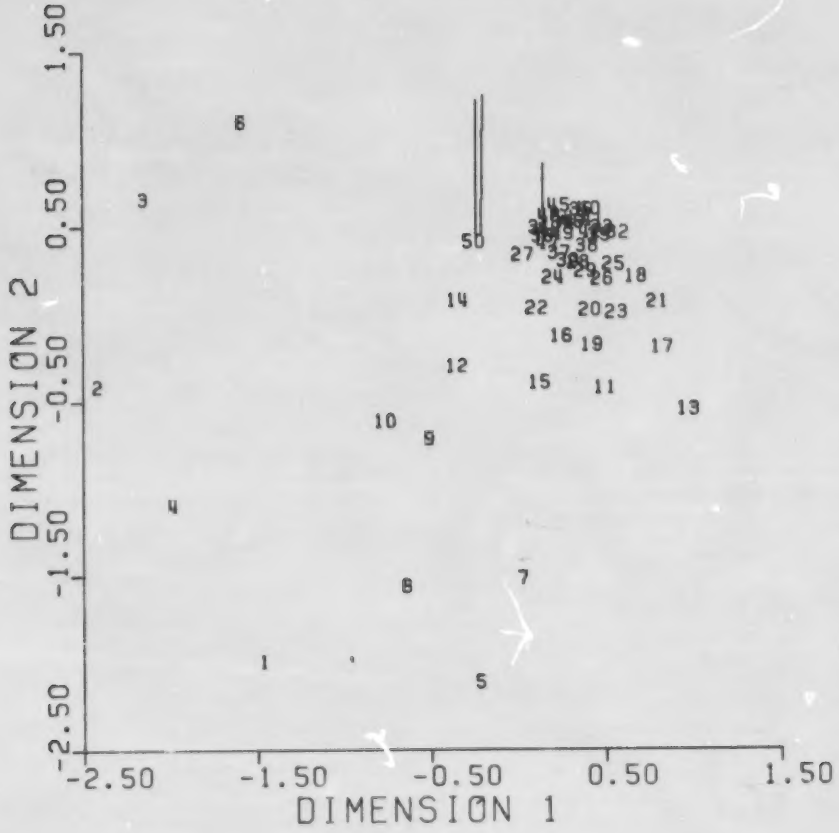
NOTE. THE ITEMS ARE NUMBERED WITH 1 THE MOST DIFFICULT AND 50 THE EASIEST

FIGURE 9

TWO-DIMENSIONAL MDSCAL SOLUTION

FOR THE ONE-FACTOR .25 GUESSING DATA

USING THE TETRACHORIC CORRELATION



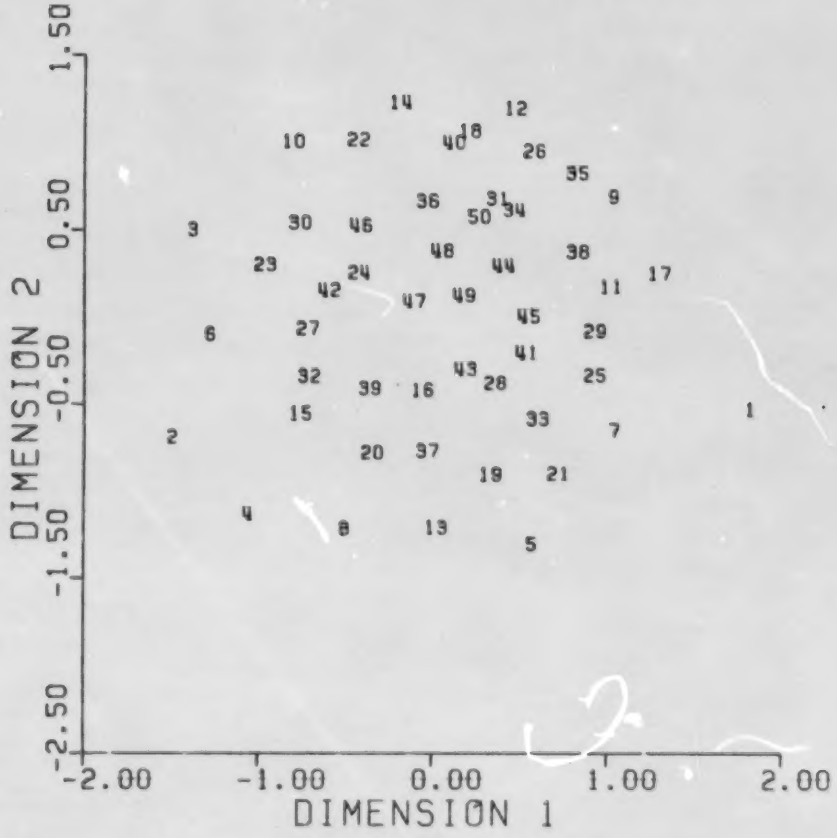
NOTE. THE ITEMS ARE NUMBERED WITH 1 THE MOST DIFFICULT AND 50 THE EASIEST

FIGURE 10

TWO DIMENSIONAL MDSCAL SOLUTION

FOR THE ONE-FACTOR .25 GUESSING DATA

USING THE ETA COEFFICIENT



NOTE. THE ITEMS ARE NUMBERED WITH 1 THE MOST DIFFICULT AND 50 THE EASIEST

solution was much higher than the solution for the no guessing data, with a value of .239, compared to the .021 obtained earlier. A plot of the results for the approval score is presented in Figure 11.

One other coefficient was considered for use with the MDSCAL procedure. That coefficient was the tetrachoric correlation corrected for guessing. To check its usefulness, the tetrachoric correlations determined from the one-dimensional, .25 guessing data were corrected for guessing using .15, .25 and .35 guessing levels. The resulting coefficients were then analyzed using the MDSCAL procedure. The results for the .15 correction gave a pattern similar to the uncorrected data, but with slightly higher stress (.234 vs. .219). The .25 correction resulted in a circular pattern similar to the no guessing data, but with hard items at one side of the plot of the solution. The stress was .320, almost as high as for the no guessing data (.334).

The .35 correction resulted in a solution with the seven hardest items in one group and all of the rest in another. The stress for this solution was a very low .074. This solution was similar to the no guessing solution for the approval score.

From the analysis of the one-dimensional, .25 guessing data, four different patterns of effects can be seen as a result of guessing. The coefficients that gave linear patterns when no guessing was present were either broadened into a wedge (agreement score) or bent into a crescent (kappa coefficient). The coefficients that gave circular patterns when no guessing was present were either stretched to one side by guessing (tetrachoric correlation) or maintained a circular pattern, but with the hard items on the outside and easy items in the middle (eta coefficient). Carroll's correction for guessing did tend to compensate for guessing effects. However, the MDSCAL solution only matched the no guessing solution if the correction matched the true guessing level. Otherwise, the solution was distorted.

Cluster Analysis As with the factor analysis and multidimensional scaling procedures, the first analysis performed with the two cluster analysis procedures was the application of the techniques to one-dimensional data with no guessing. After the no guessing analysis, the techniques were applied to the one-dimensional data set with .25 guessing to determine guessing effects. In all cases data-sets with rectangularly distributed traditional difficulty indices were used to make clearer any item difficulty effects. All of the coefficients listed previously were used for these analyses, along with both cluster analysis procedures, CLUSTER and HICLUSTER. The CLUSTER results will be presented first.

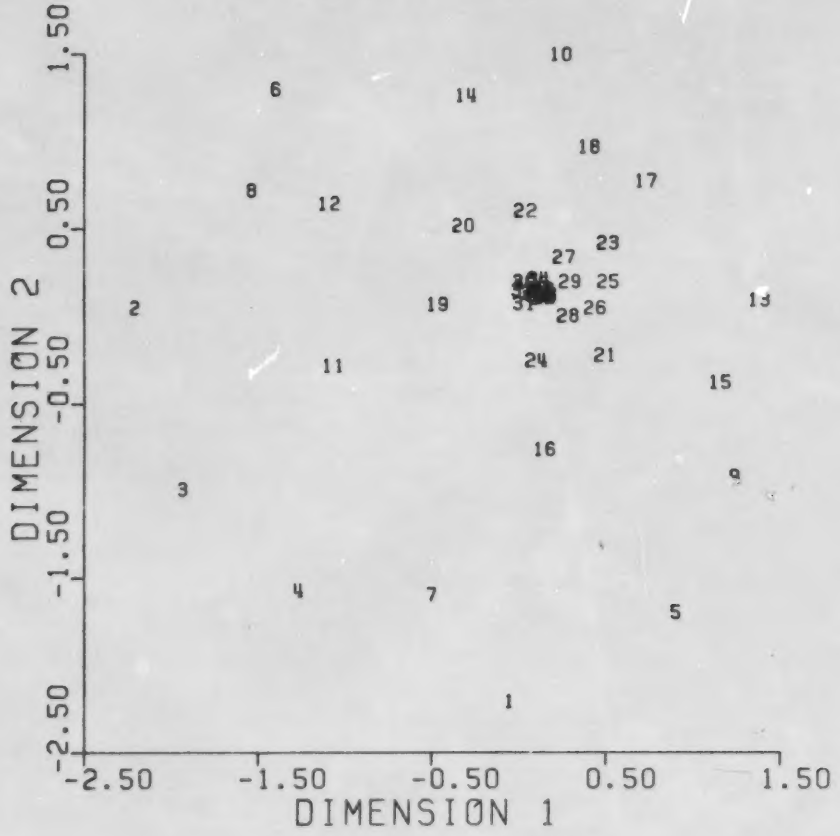
The CLUSTER results are difficult to interpret because the number of clusters obtained depends on the cutoff value used to accept an item into a cluster. Slight changes in the value result in substantial changes in the number of clusters obtained. Despite these difficulties, a pattern was determined in the results. The cluster analysis solution determined for the kappa, phi, agreement, Lijphart, koppa, tau B, and approval coefficients were all related to the difficulty of the items. That is, items of similar difficulty

FIGURE 11

TWO-DIMENSIONAL MDSCAL SOLUTION

FOR THE ONE-FACTOR .25 GUESSING DATA

USING THE APPROVAL SCORE



NOTE. THE ITEMS ARE NUMBERED WITH 1 THE MOST DIFFICULT AND 50 THE EASIEST

were clustered together. In contrast, the solutions based on Yule's Q, Yule's Y, eta, tetrachoric, phi/phi max, and gamma tended to form single large clusters or clusters unrelated to item difficulty. This result is reasonable, since the latter coefficients all yielded coefficients that are fairly independent of item difficulty, while the former set of coefficients are dependent on item difficulty. Individual results will not be presented for the CLUSTER procedure, since they are too dependent on the cutoff value for placing an item in a cluster, and no procedure is known to decide on the number of clusters.

The HICLUSTER procedure gave somewhat similar results to those of the cluster procedure. The hierarchicial solutions developed for the gamma, phi/phi max, tetrachoric, eta, Yule's Q, and Yule's Y coefficients had no discernable relationship to item difficulty, while the solution for the Lijphart, phi, kappa, agreement, tau B, koppa, and approval coefficients were related to item difficulty. Among this latter group of coefficients, three distinct patterns of entrance of the items into the clusters were noted. When using the Lijphart, koppa, and agreement coefficients the clustering procedure initially clustered the items of extreme difficulty and then worked in toward the more moderate items. The solutions based on the phi, tau B, and kappa coefficients initially clustered the middle items and then worked out toward the extremes. The approval score solution first clustered the easy items and then worked toward the most difficult. These different patterns of results reflect differences in the effect of item difficulty on the magnitude of the coefficients. Some have the highest values for middle difficulty items, while others have the highest value for items at the extremes of the difficulty range. As with the CLUSTER procedure, no procedure was known for determining the appropriate number of clusters, so no individual results will be presented here.

The application of the CLUSTER procedure to the one-dimensional, .25 guessing data gave somewhat predictable results. For the kappa, phi, agreement, Lijphart, koppa, and tau B coefficients, the clusters formed still had a tendency to be related to the item difficulty, but the relationship was not as clear. Further, more clusters were formed than when no guessing was present. This is a result of the reduced magnitude of the coefficients as a result of guessing. The results for the Yule's Q, Yule's Y, eta, tetrachoric, phi/phi max, and gamma coefficients changed somewhat from the no guessing case. The clusters formed for the guessing data had some relationship to the difficulty of the items, where none was present when guessing was not present. Correcting the tetrachoric correlations for guessing at any level did not remove this effect. The results for the approval score were very similar to those presented for the no guessing data -- the easy items formed a large cluster, while many small clusters were formed from the more difficult items.

The HICLUSTER procedure gave quite different results. The majority of the coefficients formed a hierarchical structure by grouping the easier items first, and then working down toward the hard items. The coefficients that presented this pattern were the gamma, tetrachoric, Lijphart, koppa, agreement, kappa, and approval coefficients. The Yule's Q, Yule's Y, phi, and tau B co-

efficients showed some of this effect, but the results were not as strong. The phi/phi max and eta coefficients clustered the items in essentially the same way as when guessing was not present. When the tetrachoric correlation was corrected for guessing at any level, the effects of item difficulty on the clustering was removed.

The analysis of the CLUSTER and HICLUSTER results indicate that different types of clusters are developed depending on the type of coefficient used. Some coefficients yield clusters related to item difficulty, while others do not. Guessing tends to force a relationship with item difficulty for both techniques for most of the coefficients. This will have to be taken into account when working with multi-dimensional data.

Latent Trait Analysis The analysis of the one-dimensional, no-guessing data with the LOGIST program gave exactly the results that were expected. The three-parameter logistic a-parameter estimates were all uniformly high around a value of 2.0. The b-parameter estimates were evenly spaced in the range from +3 to -3, and the c-parameters were all estimated as 0.0. These results were obtained by running the LOGIST program with the default program control values.

Similar results to that obtained for the no-guessing data were also obtained when the simulated data contained a .05, .15 or .25 guessing level, assuming multiple choice items with 4 responses. The a- and b-parameter estimates gave results similar to those described above, and the c-parameters were accurately estimated at the value used to generate the data. When the level of guessing used to generate the item data was above .25, however, the default options in the program were unable to accurately estimate the parameters. With guessing at the .35 level, the c-parameters were underestimated for all but the hard items. The a-parameter estimates tended to be low for the moderate and hard items, and the b-parameter estimates were becoming more erratic. The parameter estimates for the .25 and .35 cases are presented in Table 7. The parameter estimates are progressively worse for guessing at the .45, .55, .65, and .75 levels.

The parameter estimates obtained from the LOGIST program for the high guessing levels could be improved by releasing the constraints on the c-parameter. When the range of acceptable c-values was made larger the program did a good job of estimating the parameters at the .35 and .45 levels. Parameter estimates for higher guessing levels were still inaccurate.

In evaluating the results of these analyses, it is clear that LOGIST program does well when the guessing levels are low to moderate, and poorly when guessing is high. These results should be taken as very favorable overall, since it is unlikely that guessing on typical multiple choice items is ever as high as .65 or .75. That is, subjects with ability at $-\infty$ are unlikely to have that high a probability of obtaining a correct response to an item. When the guessing level is reasonable, the program does a very accurate job of estimating the parameters.

Table 7

Item Parameter Estimates for the One-Dimensional Data with .25 and .35 Guessing Levels and Rectangular Distribution of Difficulty

Item Number	Guessing Level					
	.25			.35		
	a	b	c	a	b	c
1	2.00	2.44	.245	.15	7.90	.250
2	2.00	2.14	.245	.08	14.57	.250
3	2.00	1.97	.240	.93	2.64	.305
4	1.91	1.73	.245	.63	2.16	.264
5	2.00	1.51	.245	1.95	1.72	.306
6	2.00	1.48	.250	1.93	1.66	.331
7	2.00	1.38	.245	1.86	1.47	.307
8	2.00	1.18	.250	1.47	1.33	.327
9	2.00	1.11	.250	.93	1.14	.250
10	2.00	1.00	.290	1.87	1.09	.328
11	2.00	.88	.245	.94	.73	.250
12	2.00	.88	.245	1.29	1.01	.334
13	2.00	.80	.250	2.00	.82	.321
14	2.00	.72	.245	2.00	.81	.367
15	2.00	.66	.240	2.00	.72	.325
16	2.00	.56	.245	1.37	.41	.250
17	1.88	.51	.250	1.30	.27	.250
18	1.91	.46	.240	2.00	.52	.310
19	2.00	.34	.245	1.41	.29	.250
20	2.00	.30	.245	1.57	.14	.250
21	2.00	.21	.245	1.70	.09	.250
22	2.00	.20	.245	1.70	.05	.250
23	2.00	.17	.250	1.47	-.00	.250
24	2.00	.08	.240	1.65	-.01	.250
25	1.95	.07	.245	1.50	-.15	.250
26	2.00	-.00	.245	1.63	-.21	.250
27	1.78	-.09	.250	1.58	-.22	.250
28	2.00	-.14	.245	1.68	-.36	.250
29	2.00	-.14	.245	2.00	-.39	.250
30	2.00	-.31	.250	2.00	-.29	.250
31	2.00	-.33	.245	2.00	-.48	.250
32	1.83	-.40	.245	2.00	-.50	.250
33	2.00	-.54	.245	2.00	-.52	.250
34	2.00	-.53	.240	2.00	-.64	.250
35	2.00	-.60	.250	2.00	-.74	.250
36	2.00	-.70	.240	1.98	-.73	.250
37	2.00	-.74	.245	2.00	-.79	.256
38	1.97	-.80	.245	2.00	-.92	.250
39	2.00	-.89	.245	2.00	-.90	.250
40	2.00	-.97	.245	1.85	-1.08	.250
41	2.00	-1.01	.245	2.00	-1.05	.250
42	2.00	-1.18	.245	2.00	-1.06	.250
43	2.00	-1.23	.245	2.00	-1.21	.250
44	2.00	-1.44	.245	2.00	-1.33	.250
45	2.00	-1.60	.245	2.00	-1.38	.250
46	2.00	-1.51	.245	2.00	-1.48	.250
47	2.00	-1.75	.245	2.00	-1.64	.250
48	2.00	-2.01	.245	2.00	-2.01	.250
49	2.00	-2.43	.245	2.00	-2.24	.250
50	2.00	-3.50	.245	1.86	-3.22	.250

Summary The purpose of this section has been to report the results of the four techniques considered in this report -- factor analysis, nonmetric multidimensional scaling, cluster analysis, and latent trait analysis--to one-dimensional data to serve as a frame of reference for the analysis of multidimensional data. The factor analysis, multidimensional scaling, and latent trait analysis gave a clear indication of the one-dimensional nature of the data when no guessing was present. When guessing was present the distorting effect could be seen in the results of each of the techniques. The precent of variance in the first factor was reduced for the factor analysis technique, along with reduced first factor loadings and the presence of extra guessing factors. The two-dimensional representations of the MDSCAL results were stretched or bent by the guessing effect, and LOGIST parameter estimates were less accurate when high guessing was present (.35 and above).

The results of the two cluster analysis procedures were harder to interpret in that it was hard to decide how many clusters were in the data. One consistent finding was that guessing was found to make the solutions more dependent on item difficulty. The problem with the determination of the number of clusters seems to make this technique less useful for forming unidimensional subsets.

Each of the above techniques was applied to two-dimensional data to determine how well the items could be sorted into unidimensional sets. Only techniques judged to perform this sorting task well were used in later analyses.

Two-Dimensional Simulated Data

The results of the application of the four techniques to the two-dimensional data will be presented in the same order as in the previous section: factor analysis, multidimensional scaling, cluster analysis, and latent trait analysis. Three two-dimensional simulated data-sets were subjected to analysis: (a) a data-set with a rectangular distribution of difficulty and no guessing; (b) a data-set with a normal distribution of difficulty and normally distributed guessing around .20; and (c) a data-set with rectangularly distributed item difficulty and constant guessing at .25. These three data sets were selected to vary the difficulty of the sorting task and the realistic nature of the data. All data-sets had 50 items, 1000 cases, and loadings for each item of .90 on one factor and .00 on the other. The factor loading matrix used to generate the data is given in Table 8.

Factor Analysis For each of the data-sets, six factor analyses were possible. These included the analyses using either the principal component or principal factor method on phi, tetrachoric, or corrected tetrachoric correlations. In some cases, maximum likelihood factor analysis was also run on the data.

The simplest of the three data-sets containing two factors had a rectangular distribution of difficulties and no guessing effect. Of the six possible analyses, the principal factor analysis on phi coefficients gave the best overall results. From this analysis, it was easy to identify the items generated from each factor, and the eigenvalues indicated two major factors and

Table 8

Factor Loading Used to Generate
the Two-Factor and Three-Factor Simulated Data

Item Number	Two-Factor		Three-Factor		
	I	II	I	II	III
1	9	0	5	-5	0
2	0	9	5	0	5
3	0	9	5	5	0
4	9	0	5	0	-5
5	9	0	5	0	5
6	0	9	5	5	0
7	0	9	5	0	5
8	9	0	5	-5	0
9	9	0	5	-5	0
10	0	9	5	-5	0
11	0	9	5	0	5
12	9	0	5	0	-5
13	0	9	5	0	5
14	9	0	5	-5	0
15	9	0	5	0	5
16	0	9	5	0	-5
17	0	9	5	0	5
18	9	0	5	5	0
19	0	9	5	0	5
20	9	0	5	0	5
21	9	0	5	0	-5
22	0	9	5	0	-5
23	0	9	5	5	0
24	9	0	5	0	5
25	0	9	5	-5	0
26	9	0	5	-5	0
27	9	0	5	0	5
28	0	9	5	-5	0
29	9	0	5	-5	0
30	0	9	5	5	0
31	0	9	5	5	0
32	9	0	5	0	-5
33	9	0	5	5	0
34	0	9	5	-5	0
35	9	0	5	-5	0
36	0	9	5	0	-5
37	0	9	5	0	5
38	9	0	5	5	0
39	0	9	5	0	5
40	9	0	5	0	-5
41	9	0	5	-5	0
42	0	9	5	5	0
43	0	9	5	0	5
44	9	0	5	0	5
45	0	9	5	-5	0
46	9	0	5	5	0
47	9	0	5	5	0
48	0	9	5	0	5
49	0	9	5	5	0
50	9	0	5	5	0

Note. All factor loadings are presented without decimal points.

and two minor were present in the data. The factor loadings resulting from this analysis are shown in Table 9. All of the other analyses performed on this data-set yielded factor loading matrices that did not clearly identify the items in each factor, or that indicated that too many factors were present.

The analysis of the data-set with rectangularly distributed difficulty and guessing set at .25 gave quite a different result. The principal component analysis of the tetrachoric correlations corrected for guessing at the .25 level gave the most accurate classification of items into the factors, but also yielded a solution with 12 eigenvalues greater than 1.0. The first two eigenvalues were clearly larger than the rest, however. Unfortunately, the results cannot usually be expected to be as good. A problem with this procedure is that the level of guessing on the test items is seldom accurately known. The principal factor approach did almost as well in correctly indicating the factor used to generate the items and showed many fewer factors present in the data (four eigenvalues greater than 1). Therefore, the principal factor approach with phi coefficients was considered the best procedure for use with this data. The factor loading matrix for the first two factors of the solution is also given in Table 9. Note the reduction in the magnitude of the factor loadings with increased guessing and with the extremity of the proportion correct on the items.

The two data-sets described above are not very realistic because tests seldom have rectangular distributions of difficulty or constant guessing. Therefore, a two-dimensional data-set with normally distributed item difficulties and normaily distributed guessing levels was also analyzed. The results of the analysis of these data were uniformly good for all of the techniques. All techniques gave information that allowed the items on each factor to be clearly identified. The only difference appeared in the number of factors indicated in the data. The principal factor analysis of phi coefficients was the only technique that accurately indicated that two factors were present. This fact, and the good showing for the other data-sets, seems to indicate that it is the technique of choice for the two-dimensional data. The results for this technique for the normally distributed data-set are also given in Table 9. Note that for all of the phi coefficient analyses reported the loadings are much lower than those used to generate the data. The results of the analysis of the tetrachoric correlations more closely approximate the magnitude of the loadings used to generate the data, but they could not be used to classify the items as accurately.

Nonmetric Multidimensional Scaling The nonmetric multidimensional scaling procedure was applied to the same three two-dimensional data-sets used with the factor analysis procedure: two dimensions, rectangular difficulty, no guessing; two dimensions, rectangular difficulty, constant .25 guessing; and two-dimensions, normal difficulty, normal .20 guessing. The results of the analysis of these data-sets will be reported in the order given above.

The MDSCAL program was run on the two-dimensional, rectangular difficulty, no guessing (SD250R.CG00) data-set using 11 similarity coefficients. The kappa and Lijphart coefficients were deleted since they give identical results to the

Table 9
Factor Loading Matrices from the Analysis
of Three Two-Dimensional Data-sets

Item	Data-set / Technique					
	SD250R.CG00		SD250R.CG25		SD250N.NG20	
	Principal	Factor/Phi	Principal	Factor/Phi	Principal	Factor/Phi
	I	II	I	II	I	II
1	15	11	-03	-00	-44	28
2	27	-14	10	05	24	40
3	36	-15	17	00	20	45
4	26	33	-04	08	-41	26
5	26	39	-05	15	-47	24
6	47	-21	22	00	27	45
7	48	-30	25	-01	25	44
8	31	49	-10	27	-43	25
9	31	52	-05	24	-48	22
10	56	-30	35	07	25	42
11	54	-35	43	07	26	40
12	34	56	-07	41	-48	25
13	58	-34	40	06	22	43
14	35	59	-03	37	-41	28
15	35	60	-06	42	-42	22
16	61	-38	46	06	23	41
17	64	-37	51	12	21	41
18	36	62	-11	49	-41	23
19	63	-39	53	11	24	41
20	37	63	-08	51	-38	24
21	43	62	-08	52	-49	20
22	64	-39	56	04	24	38
23	62	-43	53	05	25	39
24	37	64	-11	55	-49	16
25	64	-39	58	10	26	47
26	38	63	-13	58	-39	20
27	39	66	-10	54	-45	22
28	64	-41	59	04	21	43
29	38	63	-10	57	-48	20
30	65	-37	60	07	21	41
31	63	-39	58	12	26	43
32	38	59	-06	57	-41	22
33	37	60	-09	57	-39	24
34	60	-39	59	08	22	39
35	36	59	-08	62	-38	22
36	58	-36	62	10	20	42
37	58	-34	09	08	24	37
38	38	55	-06	58	-44	24
39	56	-33	58	11	25	36
40	34	56	-07	55	-44	27
41	33	49	-09	53	-42	19
42	51	-26	56	08	22	40
43	47	-28	54	08	25	34
44	25	46	-08	46	-45	23
45	42	-23	52	05	23	43
46	25	39	-08	46	-42	18
47	22	35	-06	39	-38	18
48	29	-13	40	11	25	38
49	24	-08	33	03	17	36
50	06	11	-00	17	-37	15

Note. All factor loadings are presented without decimal points.

agreement coefficient for dichotomous data. Of the remaining coefficients, those that gave a linear pattern previously gave two solutions for the two-dimensional data. The agreement coefficient yielded an oval shaped solution (See Figure 12) and the kappa, phi, and tau B coefficients resulted in configurations with the points defined by the items distributed along two roughly parallel lines (See Figure 13). Since the agreement coefficient based solution could not be used to separate the items into unidimensional sets it was dropped from further consideration. The other three "linear" coefficients could be used equally well to separate items into homogeneous sets, although the phi and tau B coefficients gave solutions with stress values smaller than that for the kappa coefficient (.061 vs. .101).

Of the six coefficients that gave a circular solution for the one-factor data, five gave a solution for the two-factor, no guessing data that sorted the items into two distinct, tight clusters. The five coefficients were: gamma, phi over phi max, Yule's Q, Yule's Y, and tetrachoric. The results for Yule's Y is presented in Figure 14. Any of these five coefficients could be used to sort the items into homogeneous sets.

The sixth "circular" coefficient was the eta coefficient. The solution obtained using this coefficient could also be used to sort the items into homogeneous sets, but the resulting plot had more spread and had a higher stress value than the previous coefficients (.200 vs. .096 and .104). Figure 15 shows a plot of this solution.

The remaining coefficient applied to this data set was the approval score. Figure 16 shows a plot of the MDSCAL solution using this similarity coefficient. It gave a butterfly shaped pattern with the easiest items in the middle. Because of the closeness of the points representing the easy items from different factors in this solution, it may not yield a result that is useful for sorting items into homogeneous sets. The stress value for the solution was .117.

When guessing was added to the parameters used to generate the two factor data, the results were only slightly different for the "linear" coefficients. The linear sets of points had somewhat greater spread for the hard items, but the two dimensions were still clearly recognizable. Figure 17 shows the results for the tau B coefficient.

The "circular" coefficients were affected somewhat more than the "linear" coefficients. The tight clusters of points found when there was no guessing effect were spread quite dramatically, showing the effect of guessing. The results for Yule's Y are shown in Figure 18, demonstrating this effect.

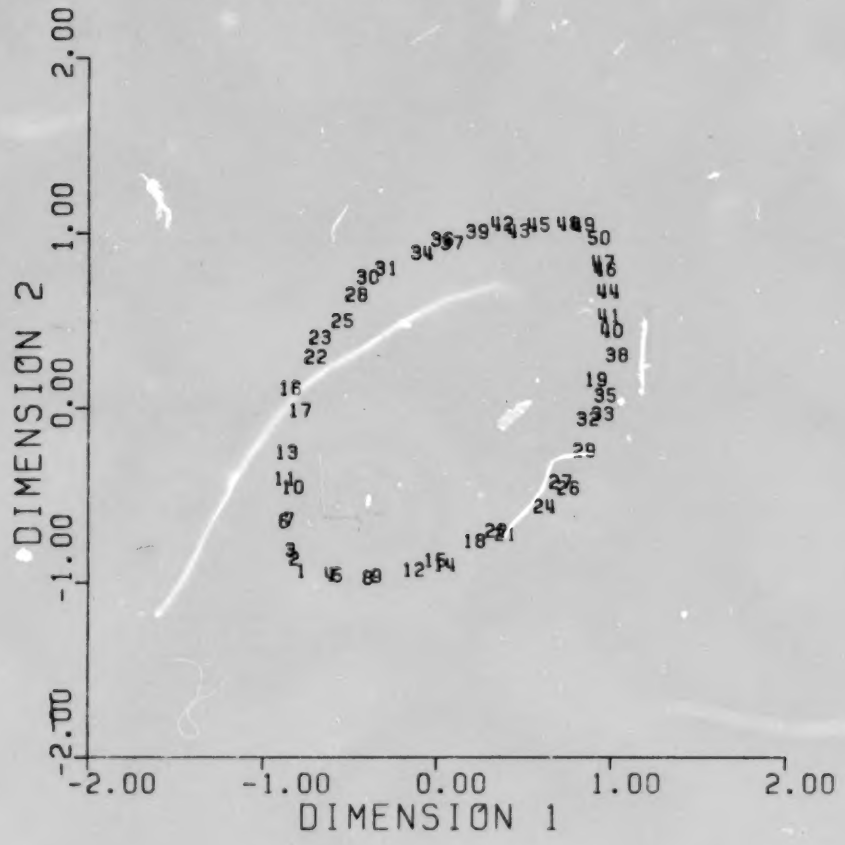
The scatter in the solutions obtained using the eta and approval coefficients increased with the presence of guessing to the point where the separate subsets of items were no longer readily identified. These two coefficients were therefore dropped from further consideration.

FIGURE 12

TWO-DIMENSIONAL MDSCAL SOLUTION

FOR THE TWO-FACTOR .00 GUESSING DATA

USING THE AGREEMENT COEFFICIENT



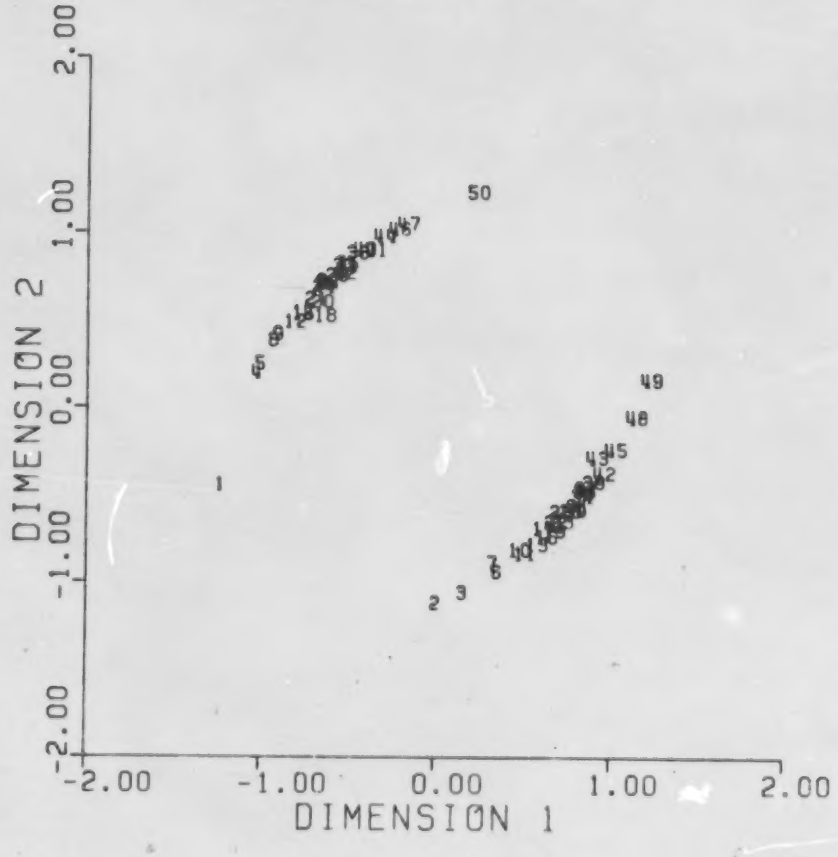
NOTE. THE ITEMS ARE NUMBERED WITH 1 THE MOST DIFFICULT AND 50 THE EASIEST

FIGURE 13

TWO-DIMENSIONAL MOSCAL SOLUTION

FOR THE TWO-FACTOR .00 GUESSING DATA

USING THE PHI COEFFICIENT



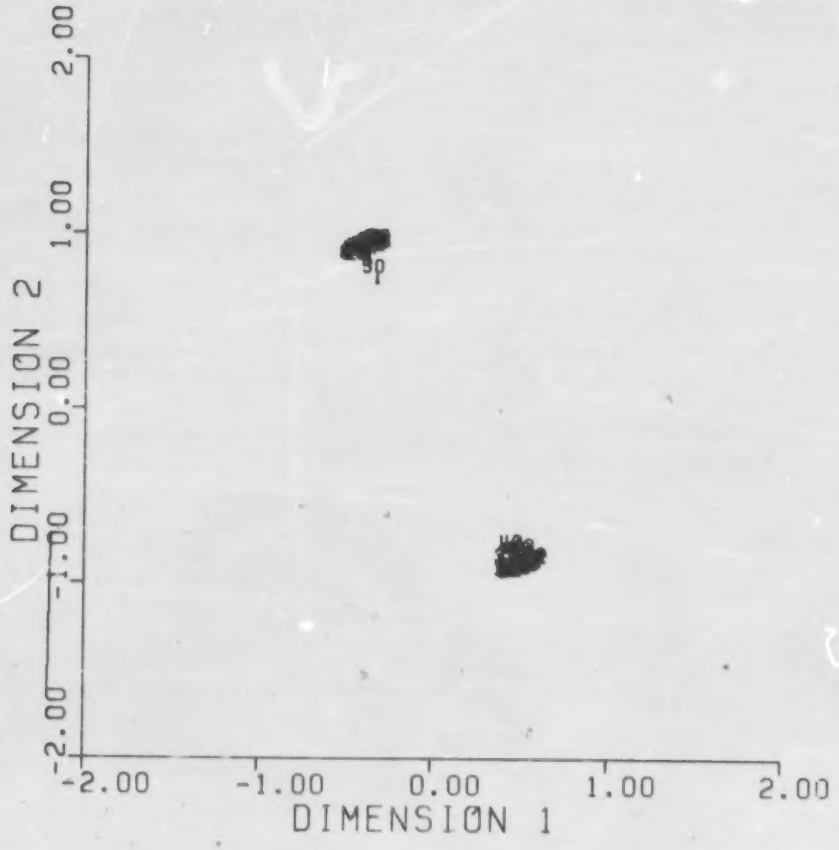
NOTE. THE ITEMS ARE NUMBERED WITH 1 THE MOST DIFFICULT AND 50 THE EASIEST

FIGURE 14

TWO-DIMENSIONAL MDSCAL SOLUTION

FOR THE TWO-FACTOR .00 GUESSING DATA

USING YULES'S Y COEFFICIENT



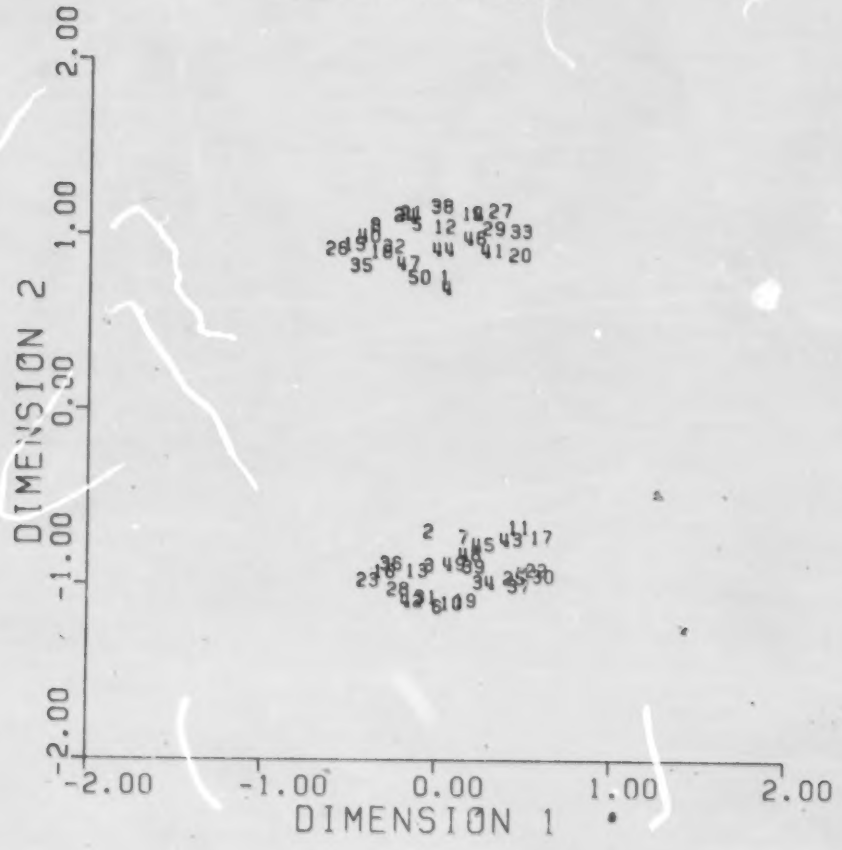
NOTE. THE ITEMS ARE NUMBERED WITH 1 THE MOST DIFFICULT AND 50 THE EASIEST

FIGURE 15

TWO-DIMENSIONAL MDSCAL SOLUTION

FOR THE TWO-FACTOR .00 GUESSING DATA

USING THE ETA COEFFICIENT



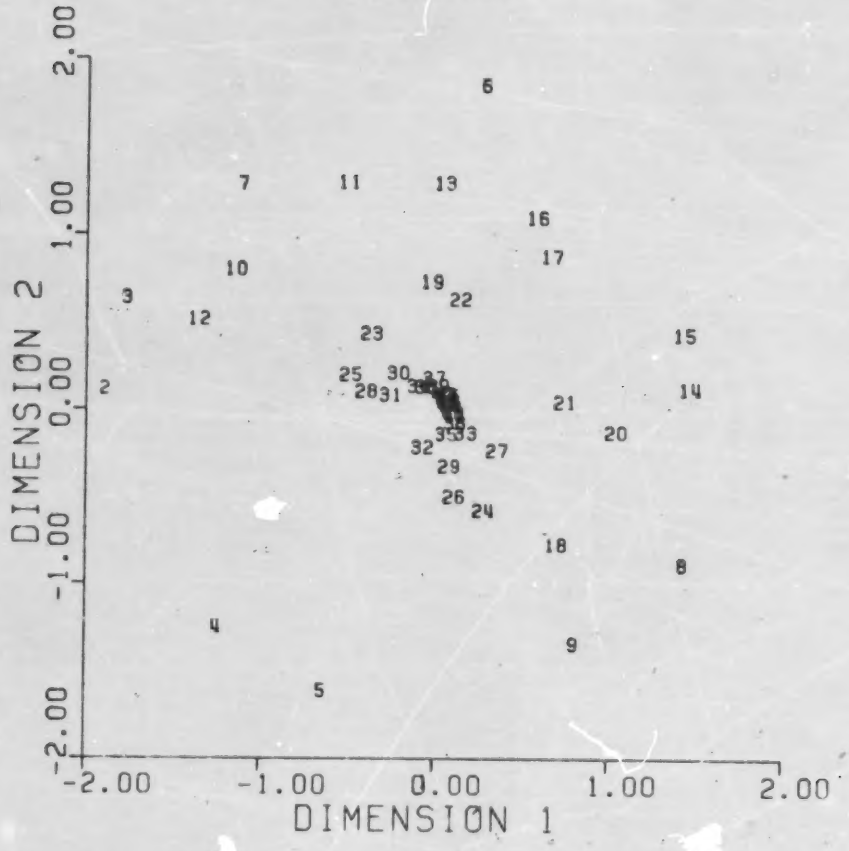
NOTE. THE ITEMS ARE NUMBERED WITH 1 THE MOST DIFFICULT AND SO THE EASIEST

FIGURE 16

TWO-DIMENSIONAL MOSCAL SOLUTION

FOR THE TWO-FACTOR .25 GUESSING DATA

USING THE APPROVAL SCORE



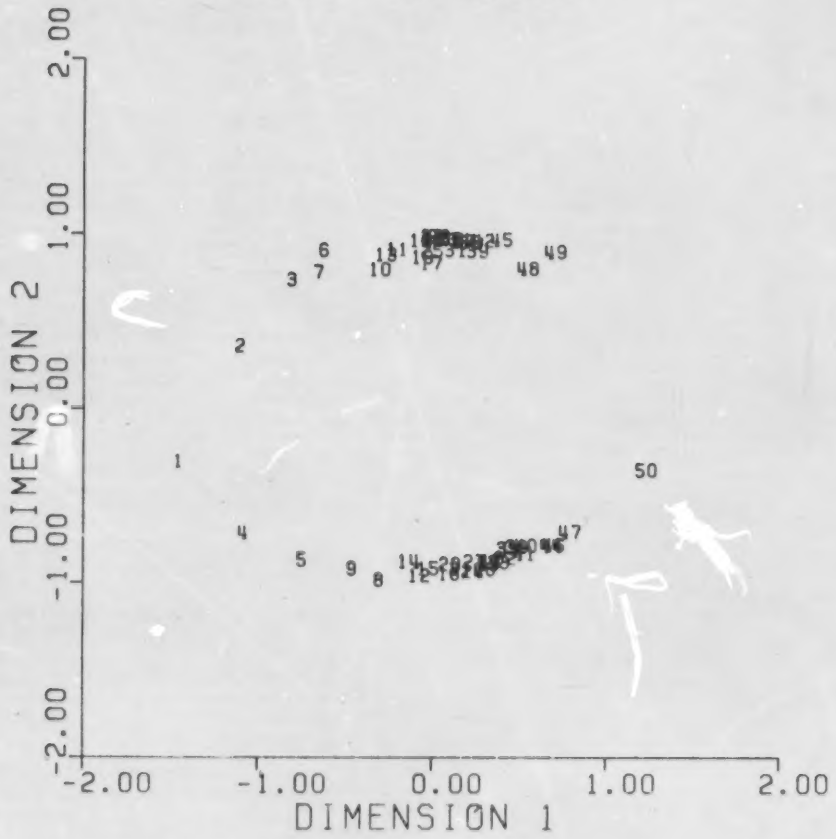
NOTE. THE ITEMS ARE NUMBERED WITH 1 THE MOST DIFFICULT AND 50 THE EASIEST

FIGURE 17

TWO-DIMENSIONAL MDSCAL SOLUTION

FOR THE TWO-FACTOR .25 GUESSING DATA

USING KENDALL'S TAU B COEFFICIENT



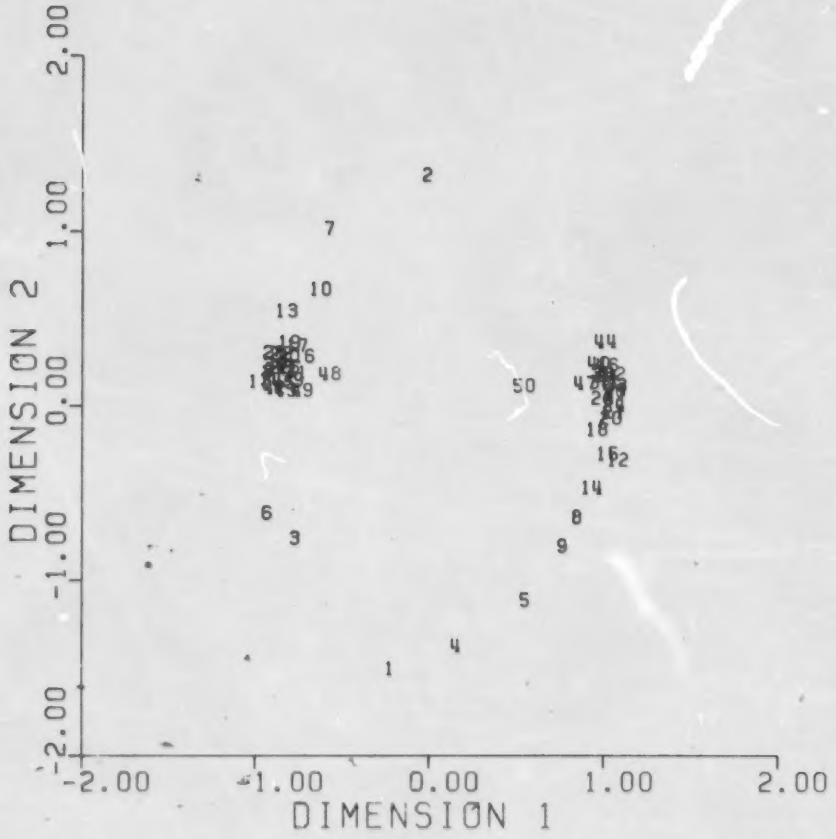
NOTE. THE ITEMS ARE NUMBERED WITH 1 THE MOST DIFFICULT AND 50 THE EASIEST

FIGURE 18

TWO-DIMENSIONAL MOSCAL SOLUTION

FOR THE TWO-FACTOR .25 GUESSING DATA

USING THE YULE'S Y CORRELATION



NOTE. THE ITEMS ARE NUMBERED WITH 1 THE MOST DIFFICULT AND 50 THE EASIEST

One final analysis was performed on this data-set. The MDSCAL program was applied to the matrix of tetrachoric correlations corrected for guessing at the .25 level. The resulting plot of the solution was somewhat clearer than that for the uncorrected tetrachoric correlations, but the stress increased from .146 to .174.

After deleting the agreement, approval, and eta coefficients from consideration because they gave ambiguous results, eight coefficients remained. These eight were computed on the two-dimensional data-set with normal difficulty and normal guessing (SD25ON.NG20). The results were uniformly good, looking approximately like Figure 14. Because of their similarity, individual results will not be presented.

The results of the analysis of the two factor data-sets show that the MDSCAL program applied to any of kappa, phi, tau B, gamma, Yule's Q, Yule's Y, tetrachoric, or corrected tetrachoric coefficients yielded solutions capable of sorting the items into the factors used to generate them. Of this set of coefficients, Kendall's tau B gave the solution with the lowest stress value.

Cluster Analysis Both the CLUSTER and HICLUSTER programs were applied to the three data-sets analyzed by the factor analysis and nonmetric multidimensional scaling procedures. The results obtained from the application of these two techniques to the data-sets were generally disappointing. While the factor analysis and multidimensional scaling procedures could accurately classify the items into the correct factor, in no case, regardless of the coefficient used, could the cluster analysis procedure do so. This poor showing occurred despite the fact that a two cluster solution was assumed in advance. If the number of clusters present in the data had not been known, the results would have been much worse, since no reasonable criterion was known for determining the number of clusters.

To demonstrate the poor quality of the information obtained from the cluster analysis procedures, the number of misclassified items based on the analyses of the data using 10 different coefficients is shown in Table 10. These results were based on a two cluster solution using the HICLUSTER procedure, and the closest to a two cluster solution that could be obtained from the CLUSTER procedure by varying the criterion for entering a cluster. The results are presented for the SD25OR.CG25 data-set. The results for the data-set with a normal distribution of item difficulties were substantially better, with few errors in classification, but the items in that data-set have been shown to be very easy to classify using the other procedures.

As can be seen from Table 10, many of the items were placed in clusters defined by items from the other factor. The agreement and approval scores yielded particularly bad results for the CLUSTER procedure because many different clusters were formed, none of which conformed to the structure used to generate the data. Ironically, the approval score, which gave the worst results for the CLUSTER program, gave the best results for the HICLUSTER program. Because of the erratic and often poor results obtained from the cluster analysis procedures,

Table 10

Errors in Classification of Items Onto Dimensions
for the CLUSTER and HICLUSTER programs
Using a Variety of Coefficients.

Coefficient	Program	
	CLUSTER*	HICLUSTER
Agreement	24	21
Approval	37	2
Eta	33	10
Gamma	7	8
Lijphart	24	21
Phi	16	22
Tau b	16	21
Tetrachoric	7	8
Yule's Q	7	8
Yule's Y	14	8

* The poor results for some of the analyses using CLUSTER were due to the fact that a two cluster solution could not be obtained.

they were removed from further consideration as item sorting techniques.

Latent Trait Analysis The application of the LOGIST program to the three two-factor data-sets gave good results for the no guessing and the normally distributed .20 guessing case, and fairly good results for the two factor data with rectangular difficulties and .25 guessing. In the former two cases, the items generated from one factor had uniformly high discrimination parameter estimates while those from the other were uniformly low. Items could be correctly classified 100% of the time. In the latter case, six iterations of the program, deleting low discriminating items after each iteration, were required in order to get a set of items that had uniformly high discrimination parameter estimates. Only one item of the 25 items retained came from the alternate factor. Unfortunately, the six iterations required about nine minutes of CPU time, compared to about 30 seconds for factor analysis. Unless the number of iterations needed to form the homogeneous item sets can be kept to a small number, this procedure may be prohibitively expensive.

Three-Dimensional Simulated Data The three dimensional data-set generated for this study was produced to match what was considered to be a reasonable model of real test data. This data-set had a general first factor with .5 loadings for each item. The second and third factors were bipolar, with half of the items having .5 loadings on one of the factors and half on the other.

The factor loading matrix used to generate the data is presented in Table 8. The .5 loadings used for this data-set were thought to be much more reasonable than the .9 loadings used for the previous data-sets.

Three procedures were applied to this data-set: factor analysis, multidimensional scaling, and latent trait analysis. As mentioned earlier, the cluster analysis procedure was dropped from consideration because it could not be used to sort items into homogeneous sets. The factor analysis results will be presented first.

Factor Analysis All six of the factor analysis solutions described for the two factor data were obtained for this data-set. These included the principal component and principal factor solutions on phi coefficients, tetrachoric correlations, and tetrachoric correlations corrected for guessing. Of these, the analysis of the tetrachoric correlations corrected for guessing clearly did not give a good representation of the structure used to generate the data. The principal factor solution could not be obtained at all, and the principal component solution did not give meaningful factors. This is probably due to the fact that the tetrachoric correlations were corrected for constant guessing at a .20 level, while the guessing level in the data varied substantially around .20. Thus, for many of the items the procedure over corrected for guessing. These results indicate that correcting for guessing is not a reasonable procedure with realistic data where the true guessing level of the items is unknown.

Of the other solutions, the principal component solution on the tetrachoric correlations, and the principal factor solution on phi coefficients gave the best results. The varimax rotation of the principal factor solution on phi coefficients was especially accurate, correctly classifying all of the items. This solution is presented in Table 11. Note that in this solution separate factors were defined by the positive and negative ends of the factors used to generate the data. The good results obtained for the analysis of phi coefficients reinforces the results obtained on the other simulated data-sets, possibly indicating that the principal factor techniques on phi coefficients should be used for item sorting with real test data.

Nonmetric Multidimensional Scaling The nonmetric multidimensional scaling analysis of the three-dimensional data using the eight coefficients selected on the basis of the previous analyses gave uniformly good results. In all cases except when the tetrachoric correlations were over corrected for guessing at the .25 level, every item could be correctly classified onto the appropriate factor. As with the factor analysis, the items from the opposite ends of the bipolar factors were put into separate clusters. Those from the same factor were at opposite ends of the diagram in a two-dimensional plot.

Although all eight coefficients could be used to accurately sort the items into homogeneous sets, there were slight differences in the stress of the solutions. Stress values ranged from .114 to .125, with Yule's Y and the tetrachoric correlations giving the smallest values. The two-dimensional MDSCAL solution for Yule's Y is given in Figure 19.

Table 11

Factor Loading Matrix from the Varimax Rotation of the Four Factor Principal Factor Solution on the Three Dimensional Data-Set.

Item	Factor			
	I	II	III	IV
1	14	51	-04	09
2	49	12	16	-06
3	21	-02	41	20
4	-05	24	22	38
5	47	11	09	00
6	23	-11	39	18
7	48	08	15	01
8	16	49	-08	16
9	16	46	-06	13
10	10	51	00	15
11	48	05	07	03
12	-04	26	13	46
13	46	09	17	-02
14	13	48	-07	07
15	54	12	07	-01
16	-07	15	05	52
17	45	16	23	-02
18	14	-03	42	14
19	45	16	16	-10
20	43	10	13	06
21	-07	14	12	41
22	-10	19	18	40
23	15	-08	51	03
24	41	16	11	-06
25	15	45	-02	10
26	12	45	-06	04
27	49	10	15	-03
28	07	47	-14	17
29	14	48	-07	14
30	21	-05	43	10
31	11	-06	46	04
32	01	13	11	48
33	16	-12	31	18
34	12	39	-02	15
35	10	43	-02	10
36	00	11	11	42
37	37	13	20	-11
38	15	-12	45	15
39	43	12	14	-03
40	-06	15	19	40
41	12	44	-02	12
42	18	-03	50	09
43	45	11	05	-01
44	48	10	07	-01
45	25	36	-17	20
46	17	-02	42	17
47	06	43	02	01
48	42	14	05	00
49	05	-04	41	06
50	16	-09	38	14

Note. All values are presented without decimal points.

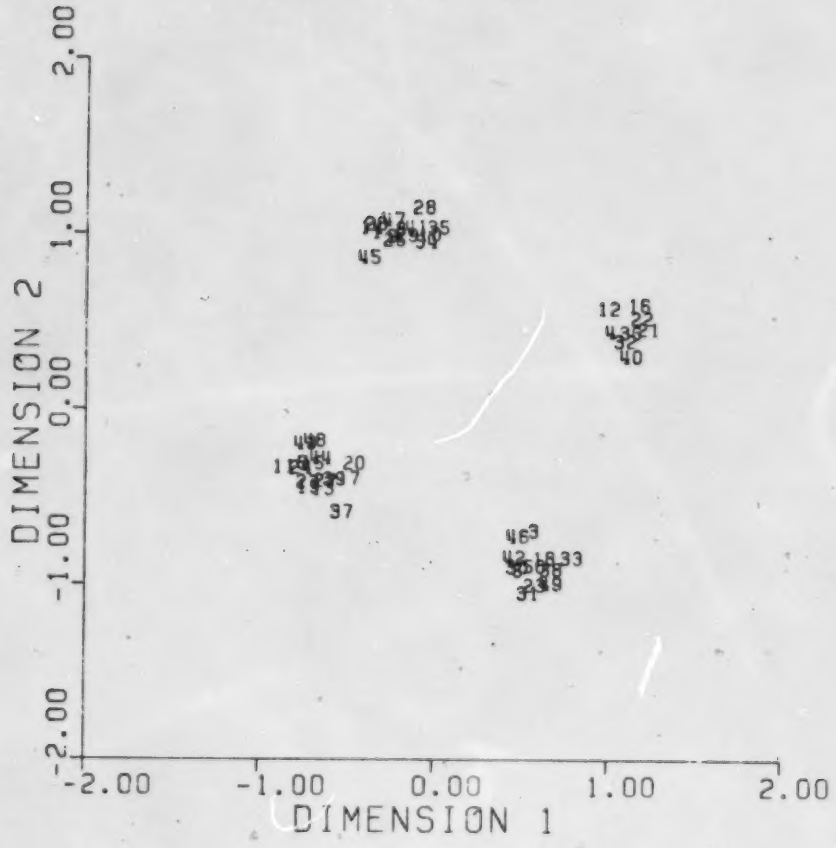
FIGURE 19

TWO-DIMENSIONAL MDSCAL SOLUTION

FOR THE THREE-FACTOR DATA SET

WITH GUESSING NORMALLY DISTRIBUTED AROUND .20

USING YULE'S Y COEFFICIENT



Latent Trait Analysis The analysis of the three factor data with the LOGIST program resulted in an accurate classification of the items onto the respective factors. The discrimination parameter estimates for the items from one end of a single bipolar factor all had uniformly high values of around 1.0, while the rest of the items had parameter estimates of .30 or less. As with the previous analyses, each end of the bipolar factors defined separate sets of items. To obtain the complete sorting of the items, three runs of the program were required.

Summary All three procedures used to analyze the three factor data-set resulted in solutions that could be used to form homogeneous item sets. The factor analysis procedures defined clear sets of items using the principal component procedure on tetrachoric correlations and the principal factor procedure on phi coefficients. The MDSCAL program gave clear solutions using the phi, tau B, kappa, tetrachoric, corrected tetrachoric, Yule's Y, Yule's Q and gamma coefficients. Only when the tetrachoric correlations were corrected at too high a level did the procedure degenerate. A similar finding was observed with the factor analysis procedures. The LOGIST analysis of the data also gave accurate sortings of the items, but three program runs were required to sort all of the items. The results of the application to a more realistic nine factor data-set will now be reported.

Nine Factor Simulated Data

The nine factor simulated data-set was the most realistic of the simulation data-sets produced. Its characteristics were designed to match those of an actual achievement test measuring nine content areas. This data-set had a general factor and eight group factors, the last one being bipolar. The major loadings on the first eight factors were all positive, reflecting the structure seen in most achievement tests. The factor loading matrix used to produce this data-set is given in Table 12.

The results of the analysis of this data-set using factor analytic techniques are similar to those obtained for the three factor data-set. Both the principal components analysis on tetrachoric correlations and the principal factor analysis on phi coefficients gave results that were easily used to sort the items into homogeneous groups. As an example of these results, the varimax rotated principal factor solution is shown in Table 13. Notice that no general factor is present in this solution. The general factor was present in the initial principal factor solution, but was rotated out with the varimax rotation.

Nonmetric Multidimensional Scaling The application of the MDSCAL program to the complex nine factor data-set using the eight coefficients selected on the basis of the previous analyses gave generally good results. Only the tetrachoric correlations corrected for guessing gave poor results. The problem with that coefficient again seemed to be over correcting for guessing due to the fact that the guessing level for individual items was unknown. The other seven coefficients gave good results, with Yule's Y, Yule's Q, gamma, and the tetrachoric correlation having slightly higher stress values than the phi, tau B and kappa coefficients,

Table 12

Factor Loading Matrix Used to Generate the Nine-Factor Data-Set.

Item	Factor								
	I	II	III	IV	V	VI	VII	VIII	IX
1	5	5	0	0	0	0	0	0	0
2	5	5	0	0	0	0	0	0	0
3	5	5	0	0	0	0	0	0	0
4	5	5	0	0	0	0	0	0	0
5	5	5	0	0	0	0	0	0	0
6	5	5	0	0	0	0	0	0	0
7	5	-068	5	0	0	0	0	0	0
8	5	-068	5	0	0	0	0	0	0
9	5	-068	5	0	0	0	0	0	0
10	5	-068	5	0	0	0	0	0	0
11	5	-068	5	0	0	0	0	0	0
12	5	-068	-064	5	0	0	0	0	0
13	5	-068	-064	5	0	0	0	0	0
14	5	-068	-064	5	0	0	0	0	0
15	5	-068	-064	5	0	0	0	0	0
16	5	-068	-064	5	0	0	0	0	0
17	5	-068	-064	-073	5	0	0	0	0
18	5	-068	-064	-073	5	0	0	0	0
19	5	-068	-064	-073	5	0	0	0	0
20	5	-068	-064	-073	5	0	0	0	0
21	5	-068	-064	-073	5	0	0	0	0
22	5	-068	-064	-073	-086	5	0	0	0
23	5	-068	-064	-073	-086	5	0	0	0
24	5	-068	-064	-073	-086	5	0	0	0
25	5	-068	-064	-073	-086	5	0	0	0
26	5	-068	-064	-073	-086	5	0	0	0
27	5	-068	-064	-073	-086	5	0	0	0
28	5	-068	-064	-073	-086	-133	5	0	0
29	5	-068	-064	-073	-086	-133	5	0	0
30	5	-068	-064	-073	-086	-133	5	0	0
31	5	-068	-064	-073	-086	-133	5	0	0
32	5	-068	-064	-073	-086	-133	5	0	0
33	5	-068	-064	-073	-086	-133	5	0	0
34	5	-068	-064	-073	-086	-133	5	0	0
35	5	-068	-064	-073	-086	-133	-176	5	0
36	5	-068	-064	-073	-086	-133	-176	5	0
37	5	-068	-064	-073	-086	-133	-176	5	0
38	5	-068	-064	-073	-086	-133	-176	5	0
39	5	-068	-064	-073	-086	-133	-176	5	0
40	5	-068	-064	-073	-086	-133	-176	-273	5
41	5	-068	-064	-073	-086	-133	-176	-273	5
42	5	-068	-064	-073	-086	-133	-176	-273	5
43	5	-068	-064	-073	-086	-133	-176	-273	5
44	5	-068	-064	-073	-086	-133	-176	-273	5
45	5	-068	-064	-073	-086	-133	-176	-273	5
46	5	-068	-064	-073	-086	-133	-176	-273	-6
47	5	-068	-064	-073	-086	-133	-176	-273	-6
48	5	-068	-064	-073	-086	-133	-176	-273	-6
49	5	-068	-064	-073	-086	-133	-176	-273	-6
50	5	-068	-064	-073	-086	-133	-176	-273	-6

Note. All values are presented without decimal points.

Table 13

Factor Loadings from the Varimax Rotation of the Principal Factor Analysis of the Nine-Factor Data-Set

Item	Factor							
	I	II	III	IV	V	VI	VII	VIII
1	-00	01	08	06	08	06	06	07
2	-04	08	02	03	07	-02	06	04
3	-05	02	02	08	03	08	05	10
4	10	08	02	00	05	04	05	09
5	02	11	08	08	03	10	09	07
6	10	03	10	06	04	06	07	02
7	04	14	10	10	13	04	06	39
8	12	03	12	11	14	10	07	37
9	07	07	06	08	-02	05	08	46
10	06	13	10	09	11	06	02	44
11	02	09	07	06	05	11	10	41
12	04	02	07	02	12	46	07	11
13	06	04	-00	08	08	45	10	09
14	13	03	09	09	03	44	00	-01
15	03	11	03	09	09	48	07	05
16	04	03	15	03	04	45	01	06
17	06	08	12	06	13	02	41	05
18	05	02	08	01	04	04	37	02
19	05	03	06	09	11	09	47	00
20	04	10	07	11	09	06	39	09
21	06	07	05	09	03	02	48	13
22	07	04	04	46	04	06	04	16
23	05	03	08	48	09	05	07	08
24	08	08	04	50	05	07	07	05
25	04	05	04	49	07	04	12	04
26	10	03	17	43	07	07	05	02
27	04	06	04	44	02	06	05	07
28	05	06	11	02	40	06	03	05
29	03	03	05	08	46	05	04	08
30	03	03	05	07	54	07	11	00
31	07	03	08	04	44	03	06	11
32	04	07	-02	06	48	05	06	03
33	02	10	06	04	45	08	12	02
34	04	06	50	05	05	02	09	02
35	05	-05	43	06	06	08	08	14
36	06	06	52	10	08	08	08	06
37	05	06	58	05	06	05	06	11
38	00	02	53	06	11	07	-01	08
39	04	07	50	08	00	09	12	01
40	04	56	09	03	07	05	04	05
41	-02	49	07	03	06	05	08	06
42	-02	59	-01	02	06	00	09	07
43	01	50	00	07	05	-02	03	06
44	-00	50	01	14	06	07	05	09
45	07	56	05	02	05	11	02	07
46	59	-01	07	07	10	07	03	05
47	63	03	07	09	04	05	04	01
48	66	02	02	06	02	07	05	07
49	75	00	04	07	03	08	03	06
50	60	03	02	06	06	04	07	06

Note. All values are presented without decimal points.

but yet yielding slightly clearer two dimensional plots. The plot for the Yule's Y coefficient is shown in Figure 20. Examination of the plot will show that the items in the data-set have been divided into nine clearly distinguished clusters.

Latent Trait Analysis The LOGIST analysis of the nine factor data-set gave disappointing results. Although the b- and c-parameter values were accurately estimated, the a-parameters gave very little indication of the items belonging to a particular factor. The a-parameter estimates varied between .41 and .81, with no noticable relationship to the factor structure. Despite the initial ambiguous look of the results, a homogeneous set of six items could be obtained by running the program eight times, deleting the items with the lowest a-value estimates after each run. Since such a procedure is clearly impractical, the LOGIST program does not seem to be a viable procedure for forming unidimensional item sets.

Summary From the analysis of the one-, two-, three-, and nine-dimensional simulated data-sets, the factor analysis and multidimensional scaling procedures seem most useful for sorting items into undimensional item sets. Of the factor analysis procedures, principal component analysis of tetrachoric correlations and principal factor analysis of phi coefficients gave the best results.

Of the seven coefficients used with the MDSCAL program, those that are not affected by item difficulty seem to give a slightly better sorting of the items than those coefficients that are affected by item difficulty. These coefficients include Yule's Y, Yule's Q, gamma, and the tetrachoric correlation. Of these, Yule's Y seems to be a good choice for forming item sets because of its ease of computation and clear solutions for the simulation data.

The results for the latent trait analysis approach indicated that the procedure is too time consuming for regular use as an item sorting procedure. Although the procedure can be used to get homogeneous item sets, it requires many separate analyses and uses substantially more computer time than the other procedures.

The cluster analysis procedures used on the simpler data-sets were found to be inadequate for item sorting. No reasonable means could be found to determine the number of clusters, and the clusters that were formed often contained only some of the items generated from the same factor.

On the basis of these results, only the principal component analysis of tetrachoric correlations, principal factor analysis of phi coefficients, and MDSCAL analysis of the seven coefficients mentioned above can be recommended for forming sets of items compatible with IRT methods. These techniques were applied to the test data from the Iowa Tests of Educational Development as a final evaluation of their capabilities.

ITED Data

The ITED data-set was produced by randomly sampling 33 items from the 69

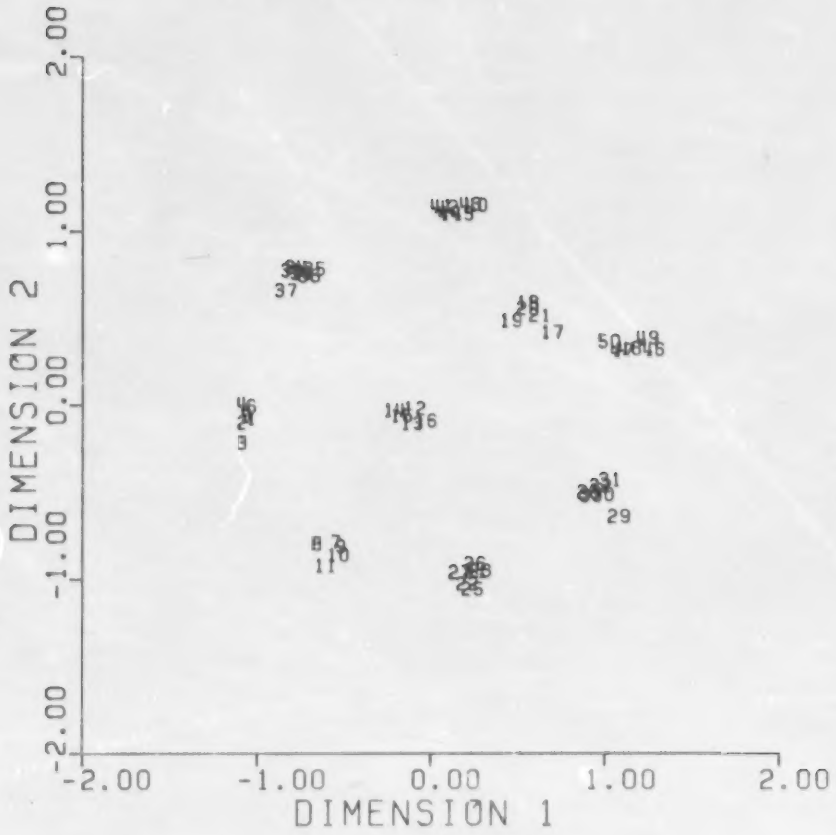
FIGURE 20

TWO-DIMENSIONAL MDSCAL SOLUTION

FOR THE NINE-FACTOR DATA SET

WITH GUESSING NORMALLY DISTRIBUTED AROUND .20

USING YULE'S Y COEFFICIENT



items in the Expression subtest and 17 items from the 36 items in the Quantitative subtest of Form Y-6 to form a 50 item test that should have had two relatively distinct components. For ease of analysis, the 33 verbal items were placed first in the test, followed by the 17 quantitative items. For these items, responses for 1000 examinees were sampled from the responses of 4000 examinees who took the test during the 1975-1976 school year. The examinees were equally divided among Grades 9, 10, 11, and 12. By producing the data-set in this way, it was hoped that a real data-set of known structure would be developed.

Factor Analysis The results of the varimax rotation of the principal factor solution of phi coefficients for the ITED data-set are presented in Table 14. The results of the principal component solution for tetrachoric correlations were similar and will not be shown. As can be seen from the table, two major factors are present in the data. Factor I is composed of most of the items from 4 to 23, which are all verbal comprehension items. Factor II is composed of most of the items from 34 to 50, which are all quantitative items. Only 17 of the 50 items in the test do not load on these two factors. Of these, six were spelling items that were mistakenly included with the verbal comprehension items (Items 28-33). These results show a relatively clear sorting of the test items into homogeneous content areas.

Nonmetric Multidimensional Scaling The results of the application of of the MDSCAL program to the inter-item similarities obtained from the seven coefficients retained up to this point were much the same, with stress values only ranging from .252 to .255 and little variation in the two-dimensional plots of the items. Figure 21 shows a representative two dimensional plot of the interrelationships of the 50 items based on Yule's Y coefficient. The initial impression obtained from this plot was that there was no clear separation of items into the different content areas. Without knowing which items were verbal and which were quantitative, this procedure could not give information for accurately sorting the items by type. Examination of higher dimension solutions also gave no clear results.

The use of knowledge concerning the content area measured by each item gives a more positive interpretation to these data. Items 34 to 50 are all quantitative items. The MDSCAL analysis resulted in a two-dimensional representation that placed all of these items in close proximity in the right side of the plot. The results of the procedure actually produced a fairly distinct separation of these items from the verbal items. Unfortunately, this pattern was very difficult to distinguish without previous knowledge of the structure of the test. For these data, at least, the factor analysis procedure gave information that was more useful for sorting items into unidimensional sets.

Discussion

The purpose of this report has been to investigate techniques for forming sets of test items that meet the assumptions of most latent trait models. That

Table 14

Varimax Factor Loading Matrix
from the Principal Factor Solution of Interitem Phi Coefficients
for the ITED Data

Item	Factor				
	I	II	III	IV	V
1	10	11	50	-00	07
2	19	13	39	07	07
3	20	12	45	10	05
4	47	18	17	11	10
5	42	21	20	15	21
6	33	23	28	07	14
7	39	18	21	24	09
8	44	23	23	14	16
9	38	23	19	06	14
10	48	23	19	14	08
11	29	24	29	04	13
12	18	17	40	09	10
13	09	16	53	13	10
14	55	25	14	11	12
15	21	21	43	23	09
16	51	23	14	13	12
17	40	20	27	08	14
18	50	29	16	12	04
19	46	20	21	10	18
20	56	18	18	16	11
21	36	19	23	10	31
22	42	20	14	15	37
23	23	22	22	12	41
24	30	22	23	08	46
25	29	17	20	16	34
26	13	21	45	14	11
27	37	16	20	24	09
28	33	26	21	36	01
29	34	25	24	38	11
30	23	22	23	62	15
31	17	06	39	17	07
32	16	15	41	28	09
33	25	26	24	45	14
34	19	39	18	21	-01
35	10	43	25	09	11
36	27	52	10	12	08
37	19	58	27	16	14
38	24	52	17	15	16
39	39	42	07	13	09
40	20	35	26	05	09
41	27	35	27	07	14
42	19	31	40	10	14
43	25	43	19	09	09
44	23	50	18	24	17
45	26	48	18	18	12
46	18	25	31	10	18
47	18	37	33	11	15
48	25	44	21	04	17
49	33	44	15	03	07
50	19	40	39	09	11

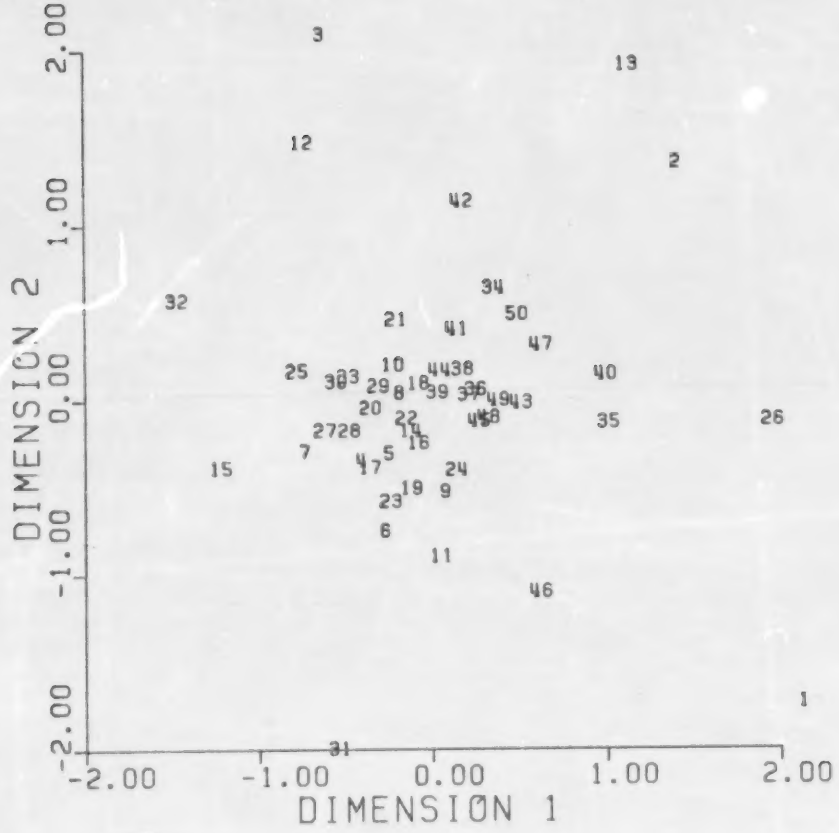
Note. All values are presented without decimal points.

FIGURE 21

TWO-DIMENSIONAL MDSCAL SOLUTION

USING THE ITED-DATA SET

USING YULE'S Y COEFFICIENT



is, procedures were evaluated for sorting items into sets that measured a single latent trait. The investigation of this problem was performed using three approaches. First, a theoretical model of guessing based on the "knowledge or random guessing" principle was produced and some theoretical results were determined. Although this model is clearly not a correct reflection of the way individuals really interact with test items, it was hoped that some insights into the effects of guessing on the observed dimensionality of item sets would be obtained.

The second approach taken in this investigation was to generate simulated test data according to the theoretical model produced in the first part of the study and use that data to evaluate factor analysis, cluster analysis, nonmetric multidimensional scaling, and latent trait analysis on their ability to form item sets measuring a single dimension. Data-sets with various numbers of factors were produced for this purpose, and the amount of guessing affecting the items was varied. Since the true structure of these data-sets was known, the quality of the results obtained from the four techniques considered was easy to evaluate.

The third approach taken in this research was to produce a data-set of known structure from existing response data on subtests of the Iowa Tests of Educational Development, and to attempt to recover that structure using the four techniques mentioned above. The data-set produced contained quantitative and verbal items, which logically should have resulted in two homogeneous subsets of items. This approach was included in the study since simulation data never really does an adequate job of modeling the interaction of examinees with test items. This "real" test data-set was the most stringent test of the procedures.

The results of the research reported here often matched what would be expected on the basis of a logical analysis of guessing and dimensionality effects, but sometimes unanticipated results were obtained. For example, the theoretical model predicted that, as guessing increased, the proportion of variance accounted for by the major factor in a test would decrease. This result was expected and was supported by the analysis of the simulation data. The review of the literature also suggested such a relationship. However, it was unexpected that an interaction would be found between the level of guessing and the saturation of an item with the major component on a test. Highly discriminating items were found to be more affected by low levels of guessing than low discriminating items, while the reverse was true for high levels of guessing (above .25). Since most multiple-choice items have average guessing levels below .25, this implies that guessing is a more serious problem for good items. This finding had not been seen in the research literature previously.

It is interesting to note that the theoretical predictions concerning guessing, including those presented in this paper, are not all consistent with each other. The results obtained by Plumlee (1952), Carroll (1945), Mattson (1965), and Denny and Remmers (1940) certainly are not consistent, and the

results presented here do not agree with any of these. This multiplicity of results reflects the complexity of the guessing phenomenon and the numerous approaches taken to modeling guessing.

The results of the analysis of the simulation data were consistent with the theoretical predictions from the models. With increased guessing, the proportion of variance accounted for by the first factor in a test decreased, and "guessing" factors appeared. Also, the magnitude of the loading of individual items decreased with increased guessing, and the effect was stronger for the more difficult items. All of these results were expected. What was not expected was that tests with rectangular distributions of traditional item difficulty were required to make these effects clearly evident. With more realistic, normal distributions of item difficulty, the guessing effects were much smaller. This suggests that guessing effects may not be too serious a problem in actual testing settings when the item difficulty is not too extreme.

The use of nonmetric multidimensional scaling, cluster analysis, and latent trait analysis had not been seen previously in the literature, so much of the results obtained was unanticipated. The two major kinds of MDSCAL plots for the one dimensional data, linear and circular, were unexpected, but further analyses showed that they were a function of the effect of item difficulty on the magnitude of the similarity coefficients. The linear plots indicate a difficulty effect, while the circular plots indicate that item difficulty has little effect.

When beginning this research, it was hoped that cluster analysis or latent trait analysis would serve as an alternative to factor analysis as a technique for purifying item pools. Unfortunately, the results of this research indicated that this hope was unjustified. Cluster analysis seems to be unsuited for this purpose. Possibly, as the research on cluster analysis progresses, better guidelines will become available for determining the number of clusters present in the test data and the procedure will, as a result, become more useful. Currently, it cannot be recommended for this use.

Latent trait analysis, the repeated application of the LOGIST program, did perform the item sorting task well, but in a very cumbersome and expensive manner. For these reasons, it cannot be recommended.

The multidimensional scaling technique applied in this research did live up to expectations. For all of the simulation data-sets the procedure presented information that could be used to identify the unifactor item sets when used with the phi, kappa, tau B, tetrachoric, gamma, Yule's Q, or Yule's Y coefficients. Unfortunately, the results were not as good for the real test data. The quantitative items were well clustered, but it was hard to distinguish between the quantitative cluster and the verbal items. Perhaps with further research better results can be obtained with real test data. The results do emphasize the fact that simulation data are not a good substitute for real data.

The procedures that performed best of all those studied were the principal component analysis of tetrachoric correlations and the principal factor analysis of phi coefficients. The interpretation of the factor analysis results was not as clear as for the MDSCAL results when simulation data were used, but was more clear for the real data. This was true even for the factor analysis of phi coefficients, which are supposed to be plagued by difficulty effects. Difficulty and guessing factors were noted when items of extreme difficulty were used, but these factors were not found for the more realistic data-sets.

The factor analysis of tetrachoric correlations worked well when the principal components technique was used, but not when the principal factor technique was used. The reason for this may be the added effect of the instability of the tetrachoric correlations when iterative estimates of communalties were made. These problems with the estimates of the tetrachoric correlations were most severe for extremely easy or difficult items.

The end result of the research reported here is that the traditional factor analysis procedure seems to perform the best of the techniques investigated for identifying items that form unidimensional item sets. The nonmetric multidimensional scaling procedure worked well for the simulation data, but the results for the real data were ambiquous. Because the study may have been biased in favor of the factor analytic procedures, due to the fact that a linear model was used to generate the simulation data, the real data-set analyses were the key to the choice of a procedure. The best procedure when using this data-set was the factor analysis procedure.

Summary and Conclusions

The effects of guessing on techniques for sorting items into sets that measure the same single dimension were determined using theoretical, simulation, and real data analyses. The theoretical results showed that, as guessing increased, the percent of variance accounted for by the major component of a test decreased. Guessing was also found to affect highly discriminating items more than low discriminating items. The results of the theoretical analyses presented here did not match those presented previously in the literature.

The factor analysis of simulated data-sets mirrored the theoretical results. As guessing increased, the proportion of variance accounted for by the first factor of the test declined. The magnitude of the loading of the items on the factor was also reduced. This effect was strongest for the most difficult items. When items of extreme difficulty were present in the simulated tests, guessing and difficulty factors were found.

The application of the nonmetric multidimensional scaling procedure to the simulation data gave different results depending on the similarity coefficient that was used. If the coefficient were affected by the difficulty of the items,

a linear pattern was found when the solution was plotted. If the similarity coefficient were not affected by the difficulty of the items, a circular plot was obtained. Guessing distorted these patterns, but the MDSCAL procedure still separated the items into homogeneous sets with little difficulty when simulation data were used. The procedure did not give adequate results for real data.

Cluster analysis and latent trait analysis were not found to be useful for sorting items into unidimensional sets. The cluster analysis procedure tended to give too many small clusters, and no way was known for combining them into larger clusters that corresponded to the known structure of the data. Repeated application of the LOGIST program to sort the items into undimensional sets worked, but was too expensive and cumbersome.

Of the procedures studied, the principal component analysis of tetrachoric correlations and the principal factor analysis of phi coefficients gave the most consistently positive results. Until a better method can be found, these time honored procedures should continue to be used to form unidimensional tests.

REFERENCES

Barr, Goodnight, Sall, and Helwig Statistical analysis system. Raleigh, NC: SAS Institute, Inc., 1976.

Carroll, J. B. The effect of difficulty and chance success on correlations between items or between tests. Psychometrika, 1945, 10(1), 1-19.

Christofferson, A. Factor analysis of dichotomized variables. Psychometrika, 1975, 40(1), 5-32.

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 1968, 70(4), 213-220.

Computing Services Office. CSO Volume 9 (User 9): Statistical Systems-SOUPAC Program Descriptions. Urbana, Illinois: University of Illinois at Urbana-Champaign, 1974.

Denny, H. R. and Remmers, H. H. Reliability of multiple-choice measuring instruments as a function of the Spearman-Brown prophecy formula, II. Journal of Educational Psychology, 1940, 31, 699-704.

Divgi, D. Dimensionality of binary items: Use of mixed models. Paper presented at the meeting of the National Council on Measurement in Education, Boston, April, 1980.

Hays, W. L. Statistics for psychologists. New York: Holt, 1963.

International Mathematical and Statistical Libraries, Inc. The IMSL Library Reference Manual, Houston, 1979.

Institute for Social Research. Osiris III. Ann Arbor: University of Michigan, 1974.

Iowa Testing Programs. Iowa Tests of Educational Development Form Y-6, Iowa City, IA, 1972.

Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. Psychometrika, 1964, 29, 115-129.

Lord, F. M., and Novick, M. R. Statistical theories of mental test scores. New York: Audison-Wesley, 1968.

MacRae, D. Issues and parties in legislative voting. New York: Harper and Row, 1970.

Mattson, D. The effects of guessing on the standard error of measurement and the reliability of test scores. Educational and Psychological Measurement, 1965, 25(3), 727-730.

Nie, Hull, Jenkins, Steinbrenner, and Brent. Statistical package for the social sciences. New York: McGraw-Hill Book Company, 1975.

Plumlee, L. B. The predicted and observed effect of chance success on multiple-choice test validity. Psychometrika, 1954, 19(1), 65-70.

Reckase, M. D. Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 1979, 4(3), 207-230.

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function, I and II. Psychometrika, 1962, 27, 125-140.

Weisberg, H. F. Dimensional analysis of legislative roll calls. Unpublished doctoral dissertation, University of Michigan, 1968.

Wherry, Sr., R. J., Naylor, J. C., Wherry, Jr., R. J., and Fallis, R. F. Generating multiple samples of multivariate data with arbitrary population parameters. Psychometrika, 1965, 30, 303-313.

Wood, R. L., Wingersky, M. S., and Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (Research Memorandum RM-76-6). Princeton, NJ: Educational Testing Service, June, 1976.

APPENDIX A

Similarity Coefficients

Many of the coefficients used in this study are based on the responses of two items as summarized in a 2x2 or 3x3 contingency table. For consistency the first 10 coefficients will be described using the following 2x2 table arrangement:

Item i \ Item j	0	1	
0	a	b	a+b
1	c	d	c+d
	a+c	b+d	N=a+b+c+d

where a, b, c, and d are cell frequencies and N is the total number of examinees.

Agreement Coefficient

The agreement coefficient (Weisberg, 1968) is the proportion of examinees responding in the same way to both items, and is given by:

$$C_1 = \frac{a+d}{N}.$$

Approval Score

The approval score (Weisberg, 1968) is the proportion of examinees passing both items, and is given by:

$$C_2 = \frac{d}{N}.$$

Eta Coefficient

The eta coefficient (Weisberg, 1968) is a measure of any type of association between two variables. It is usually used to determine the association between a nominal variable and an interval variable. This eta coefficient is in no way

related to the eta coefficient used in analysis of variance procedures. This eta is given by the following formulae:

1) if $ad > bc$ and $b > c$,

$$\eta = \frac{(ad - bc)(b - c)}{(c + d)(c + a)(b + c)};$$

2) if $ad > bc$ and $b < c$

$$\eta = \frac{(ad - bc)(c - b)}{(b + d)(b + a)(c + b)};$$

3) if $ad < bc$ and $d > a$,

$$\eta = \frac{(ad - bc)(d - a)}{(a + c)(a + b)(a + d)};$$

and 4) if $ad < bc$ and $d < a$,

$$\eta = \frac{(ad - bc)(a - d)}{(d + c)(d + b)(a + d)}.$$

Kappa Coefficient

The kappa coefficient (Cohen, 1968) is essentially an agreement score corrected for chance agreement, and is given by:

$$k = \frac{(a + d) - [(a + b)(a + c) + (c + d)(b + d)]}{1 - [(a + b)(a + c) + (c + d)(b + d)]}.$$

Koppa Coefficient

The koppa coefficient (MacRae, 1970) is the agreement score corrected for disagreement, and is given by:

$$k = \frac{a + d - (b + c)}{N}.$$

Phi Coefficient

The phi coefficient (MacRae, 1970) is a Pearson product moment correlation between binary variables, and is given by:

$$\phi = \frac{ad - bc}{\sqrt{(a + c)(b + d)(a + b)(c + d)}}.$$

Phi/Phimax Coefficient

The phi/phimax coefficient (Weisberg, 1968) is the phi coefficient divided by the maximum possible phi coefficient that could be obtained from a table with the same marginals. This procedure corrects the phi coefficient for item difficulty effects. The phi/phimax coefficient is given by the following formulae:

1) if $ad > bc$ and $b < c$,

$$\phi' = \frac{ad - bc}{(b + d)(b + a)};$$

2) if $ad > bc$ and $b > c$,

$$\phi' = \frac{ad - bc}{(c + d)(c + a)};$$

3) if $ad < bc$ and $d > a$,

$$\phi' = \frac{ad - bc}{(a + c)(a + b)};$$

and 4) if $ad < bc$ and $d < a$,

$$\phi' = \frac{ad - bc}{(d + c)(d + b)}.$$

Tetrachoric Correlation

The tetrachoric correlation (MacRae, 1970) is an estimate of the correlation between two continuous variables having a bivariate normal distribution. It has been assumed that the variables have been artifically dichotomized to produce the 2x2 table obtained for the two items. The tetrachoric correlation is approximated by:

$$r_t = \sin\left(\frac{\pi}{2} \cdot \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}\right),$$

where the value in the parenthesis is in radians. The tetrachoric correlations were corrected for guessing from the 2x2 table using the procedure set out by Carroll (1945).

Yule's Q Coefficient

Yule's Q (MaeRae, 1970) is a measure of the power of one variable to predict another, and is given by:

$$Q = \frac{ad - bc}{ad + bc}.$$

Yule's Q is a special case of Goodman and Kruskal's gamma coefficient.

Yule's Y Coefficient

Yule's Y (MaeRae, 1970) is given by:

$$Y = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}.$$

It can be seen that r_t, Q, and Y are transformations of each other.

The remaining four coefficients are best described using the following table:

		Item j			
		0	1	2	
item i	0	a	b	c	a + b + c
	1	d	e	f	d + e + f
	2	g	h	i	g + h + i
		a+d+g	b+e+h	c+f+i	N

where zero represents failing the item, 2 represents passing the item, and 1 represents a neutral or intermediate response.

Goodman and Kruskal's Gamma Coefficient

Goodman and Kruskal's gamma (Hays, 1963) is given by:

$$r = \frac{S_1 - S_2}{S_1 + S_2}$$

where

$$S_1 = a\,(e+f+h+i) + b(f+i) + d(h+i) + ei$$

and

$$S_2 = c\,(d+e+tg+h) + b(d+g) + f(g+h) + eg.$$

The gamma coefficient was developed as a measure of association between ordinal variables.

Kendall's tau B Coefficient

Kendall's tau B (Hays, 1963) is given by:

$$t_b = \frac{S_1 - S_2}{\sqrt{S_3}}$$

where S_1 and S_2 are as set out above, and

$$S_3 = [S_1 + S_2 + a(b+c) + bc + d(e+f) + ef + g(h+i) + hi] \times [S_1 + S_2 + a(d+g) + dg + b(e+h) + eh + c(f+i) + fi].$$

Lijphart's Index

Lijphart's index (MaeRae, 1970) was developed as a measure of voter agreement, and is given by:

$$i = \frac{A + \frac{d + b + h + f}{2}}{N}$$

where A is the agreement score.

Pearson's Correlation

This coefficient is the traditional product moment correlation coefficient, and is given by:

$$r = \frac{N(a+i-c-g) - (T1-T2)(T3-T4)}{\sqrt{D1D2}}$$

where

$$T1 = g + h + i,$$
$$T2 = a + b + c,$$
$$T3 = c + f + i,$$
$$T4 = a + g + d,$$
$$D1 = N(T1+T2) - (T1-T2)^2,$$

and

$$D2 = N(T3+T4) - (T3-T4)^2.$$

Navy

1 Dr. Jack R. Borsting
Provost & Academic Dean
U.S. Naval Postgraduate School
Monterey, CA 93940

1 Dr. Robert Breaux
Code N-711
NAVTRAEQUIPCEN
Orlando, FL 32813

1 Chief of Naval Education and Training
Liason Office
Air Force Human Resource Laboratory
Flying Training Division
WILLIAMS AFB, AZ 85224

1 CDR Mike Curran
Office of Naval Research
800 N. Quincy St.
Code 270
Arlington, VA 22217

1 Dr. Richard Elster
Department of Administrative Sciences
Naval Postgraduate School
Monterey, CA 93940

1 DR. PAT FEDERICO
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

1 Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. John Ford
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Henry M. Halff
Department of Psychology,C-009
University of California at San Diego
La Jolla, CA 92093

Navy

1 Dr. Patrick R. Harrison
Psychology Course Director
LEADERSHIP & LAW DEPT. (7b)
DIV. OF PROFESSIONAL DEVELOPMMENT
U.S. NAVAL ACADEMY
ANNAPOLIS, MD 21402

1 CDR Charles W. Hutchins
Naval Air Systems Command Hq
AIR-340F
Navy Department
Washington, DC 20361

1 CDR Robert S. Kennedy
Head, Human Performance Sciences
Naval Aerospace Medical Research Lab
Box 29407
New Orleans, LA 70189

1 Dr. Norman J. Kerr
Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington, TN 38054

1 Dr. William L. Maloy
Principal Civilian Advisor for
Education and Training
Naval Training Command, Code 00A
Pensacola, FL 32508

1 Dr. Kneale Marshall
Scientific Advisor to DCNO(MPT)
OP01T
Washington DC 20370

1 CAPT Richard L. Martin, USN
Prospective Commanding Officer
USS Carl Vinson (CVN-70)
Newport News Shipbuilding and Drydock Co
Newport News, VA 23607

1 Dr. James McBride
Navy Personnel R&D Center
San Diego, CA 92152

1 Ted M. I. Yellen
Technical Information Office, Code 201
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

Navy

1 Library, Code P201L
Navy Personnel R&D Center
San Diego, CA 92152

6 Commanding Officer
Naval Research Laboratory
Code 2627
Washington, DC 20390

1 Psychologist
ONR Branch Office
Bldg 114, Section D
666 Summer Street
Boston, MA 02210

1 Psychologist
ONR Branch Office
536 S. Clark Street
Chicago, IL 60605

1 Office of Naval Research
Code 437
800 N. Quincy SStreet
Arlington, VA 22217

5 Personnel & Training Research Programs
(Code 458)
Office of Naval Research
Arlington, VA 22217

1 Psychologist
ONR Branch Office
1030 East Green Street
Pasadena, CA 91101

1 Office of the Chief of Naval Operations
Research Development & Studies Branch
(OP-115)
Washington, DC 20350

1 LT Frank C. Petho, MSC, USN (Ph.D)
Selection and Training Research Division
Human Performance Sciences Dept.
Naval Aerospace Medical Research Laborat
Pensacola, FL 32508

1 Dr. Bernard Rimland (03B)
Navy Personnel R&D Center
San Diego, CA 92152

Navy

1 Dr. Worth Scanland, Director
Research, Development, Test & Evaluation
N-5
Naval Education and Training Command
NAS, Pensacola, FL 32508

1 Dr. Robert G. Smith
Office of Chief of Naval Operations
OP-987H
Washington, DC 20350

1 Dr. Alfred F. Smode
Training Analysis & Evaluation Group
(TAEG)
Dept. of the Navy
Orlando, FL 32813

1 Dr. Richard Sorensen
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Ronald Weitzman
Code 54 WZ
Department of Administrative Sciences
U. S. Naval Postgraduate School
Monterey, CA 93940

1 Dr. Robert Wisher
Code 309
Navy Personnel R&D Center
San Diego, CA 92152

1 DR. MARTIN F. WISKOFF
NAVY PERSONNEL R& D CENTER
SAN DIEGO, CA 92152

Army

1 Technical Director
U. S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Myron Fischl
U.S. Army Research Institute for the Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Dexter Fletcher
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria,VA 22333

1 Dr. Michael Kaplan
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

1 Dr. Milton S. Katz
Training Technical Area
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Harold F. O'Neil, Jr.
Attn: PERI-OK
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 DR. JAMES L. RANEY
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

1 Mr. Robert Ross
U.S. Army Research Institute for the Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

Army

1 Dr. Robert Sasmor
U. S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Commandant
US Army Institute of Administration
Attn: Dr. Sherrill
FT Benjamin Harrison, IN 46256

1 Dr. Frederick Steinheiser
Dept. of Navy
Chief of Naval Operations
OP-113
Washington, DC 20350

1 Dr. Joseph Ward
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Air Force

1 Air Force Human Resources Lab
AFHRL/MPD
Brooks AFB, TX 78235

1 Dr. Earl A. Alluisi
HQ, AFHRL (AFSC)
Brooks AFB, TX 78235

1 Research and Measurment Division
Research Branch, AFMPC/MPCYPR
Randolph AFB, TX 78148

1 Dr. Malcolm Ree
AFHRL/MP
Brooks AFB, TX 78235

1 Dr. Marty Rockway
Technical Director
AFHRL(OT)
Williams AFB, AZ 58224

Marines

1 H. William Greenup
Education Advisor (E031)
Education Center, MCDEC
Quantico, VA 22134

1 Director, Office of Manpower Utilization
HQ, Marine Corps (MPU)
BCB, Bldg. 2009
Quantico, VA 22134

1 Major Michael L. Patrow, USMC
Headquarters, Marine Corps
(Code MPI-20)
Washington, DC 20380

1 DR. A.L. SLAFKOSKY
SCIENTIFIC ADVISOR (CODE RD-1)
HQ, U.S. MARINE CORPS
WASHINGTON, DC 20380

CoastGuard

1 Mr. Thomas A. Warm
U. S. Coast Guard Institute
P. O. Substation 18
Oklahoma City, OK 73169

Other DoD

12 Defense Technical Information Center
Cameron Station, Bldg 5
Alexandria, VA 22314
Attn: TC

1 Dr. William Graham
Testing Directorate
MEPCOM/MEPCT-P
Ft. Sheridan, IL 60037

1 Military Assistant for Training and
Personnel Technology
Office of the Under Secretary of Defense
for Research & Engineering
Room 3D129, The Pentagon
Washington, DC 20301

1 Dr. Wayne Sellman
Office of the Assistant Secretary
of Defense (MRA & L)
2B269 The Pentagon
Washington, DC 20301

1 DARPA
1400 Wilson Blvd.
Arlington, VA 22209

Civil Govt

1 Dr. Andrew R. Molnar
Science Education Dev.
and Research
National Science Foundation
Washington, DC 20550

1 Dr. Vern W. Urry
Personnel R&D Center
Office of Personnel Management
1900 E Street NW
Washington, DC 20415

1 Dr. Joseph L. Young, Director
Memory & Cognitive Processes
National Science Foundation
Washington, DC 20550

Non Govt

1 Dr. Erling B. Andersen
Department of Statistics
Studiestraede 6
1455 Copenhagen
DENMARK

1 1 psychological research unit
Dept. of Defense (Army Office)
Campbell Park Offices
Canberra ACT 2600, Australia

1 Dr. Isaac Bejar
Educational Testing Service
Princeton, NJ 08450

1 Capt. J. Jean Belanger
Training Development Division
Canadian Forces Training System
CFTSHQ, CFB Trenton
Astra, Ontario KOK 1B0

1 CDR Robert J. Biersner
Program Manager
Human Performance
Navy Medical R&D Command
Bethesda, MD 20014

1 Dr. Menucha Birenbaum
School of Education
Tel Aviv University
Tel Aviv, Ramat Aviv 69978
Israel

1 Dr. Werner Birke
DezWPs im Streitkraefteamt
Postfach 20 50 03
D-5300 Bonn 2
WEST GERMANY

1 Liaison Scientists
Office of Naval Research,
Branch Office , London
Box 39 FPO New York 09510

1 Col Ray Bowles
800 N. Quincy St.
Room 804
Arlington, VA 22217

Non Govt

1 Dr. Robert Brennan
American College Testing Programs
P. O. Box 168
Iowa City, IA 52240

1 DR. C. VICTOR BUNDERSON
WICAT INC.
UNIVERSITY PLAZA, SUITE 10
1160 SO. STATE ST.
OREM, UT 84057

1 Dr. John B. Carroll
Psychometric Lab
Univ. of No. Carolina
Davie Hall 013A
Chapel Hill, NC 27514

1 Charles Myers Library
Livingstone House
Livingstone Road
Stratford
London E15 2LJ
ENGLAND

1 Dr. Kenneth E. Clark
College of Arts & Sciences
University of Rochester
River Campus Station
Rochester, NY 14627

1 Dr. Norman Cliff
Dept. of Psychology
Univ. of So. California
University Park
Los Angeles, CA 90007

1 Dr. William E. Coffman
Director, Iowa Testing Programs
334 Lindquist Center
University of Iowa
Iowa City, IA 52242

1 Dr. Meredith P. Crawford
American Psychological Association
1200 17th Street, N.W.
Washington, DC 20036

Non Govt

1 Dr.,Fritz Drasgow
Yale School of Organization and Manageme
Yale University
Box 1A
New Haven, CT 06520

1 Dr. Mavin D. Dunnette
Personnel Decisions Research Institute
2415 Foshay Tower
821 Marguette Avenue
Mineapolis, MN 55402

1 Mike Durmeyer
Instructional Program Development
Building 90
NET-PDCD
Great Lakes NTC, IL 60088

1 ERIC Facility-Acquisitions
4833 Rugby Avenue
Bethesda, MD 20014

1 Dr. Benjamin A. Fairbank, Jr.
McFann-Gray & Associates, Inc.
5825 Callaghan
Suite 225
San Antonio, Texas 78228

1 Dr. Leonard Feldt
Lindquist Center for Measurment
University of Iowa
Iowa City, IA 52242

1 Dr. Richard L. Ferguson
The American College Testing Program
P.O. Box 168
Iowa City, IA 52240

1 Dr. Victor Fields
Dept. of Psychology
Montgomery College
Rockville, MD 20850

1 Univ. Prof. Dr. Gerhard Fischer
Liebiggasse 5/3
A 1010 Vienna
AUSTRIA

Non Govt

1 Professor Donald Fitzgerald
University of New England
Armidale, New South Wales 2351
AUSTRALIA

1 Dr. Edwin A. Fleishman
Advanced Research Resources Organ.
Suite 900
4330 East West Highway
Washington, DC 20014

1 Dr. John R. Frederiksen
Bolt Beranek & Newman
50 Moulton Street
Cambridge, MA 02138

1 DR. ROBERT GLASER
LRDC
UNIVERSITY OF PITTSBURGH
3939 O'HARA STREET
PITTSBURGH, PA 15213

1 Dr. Bert Green
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

1 Dr. Ron Hambleton
School of Education
University of Massechusetts
Amherst, MA 01002

1 Dr. Chester Harris
School of Education
University of California
Santa Barbara, CA 93106

1 Dr. Lloyd Humphreys
Department of Psychology
University of Illinois
Champaign, IL 61820

1 Library
HumRRO/Western Division
27857 Berwick Drive
Carmel, CA 93921

Non Govt

1 Dr. Steven Hunka
Department of Education
University of Alberta
Edmonton, Alberta
CANADA

1 Dr. Earl Hunt
Dept. of Psychology
University of Washington
Seattle, WA 98105

1 Dr. Huynh Huynh
College of Education
University of South Carolina
Columbia, SC 29208

1 Professor John A. Keats
University of Newcastle
AUSTRALIA 2308

1 Mr. Marlin Kroger
1117 Via Goleta
Palos Verdes Estates, CA 90274

1 Dr. Michael Levine
Department of Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

1 Dr. Charles Lewis
Faculteit Sociale Wetenschappen
Rijksuniversiteit Groningen
Oude Boteringestraat 23
9712GC Groningen
Netherlands

1 Dr. Robert Linn
College of Education
University of Illinois
Urbana, IL 61801

1 Dr. Frederick M. Lord
Educational Testing Servic
Princeton, NJ 08540

1 Dr. Gary Marco
Educational Testing Service
Princeton, NJ 08450

Non Govt

1 Dr. Scott Maxwell
Department of Psychology
University of Houston
Houston, TX 77004

1 Dr. Samuel T. Mayo
Loyola University of Chicago
820 North Michigan Avenue
Chicago, IL 60611

1 Professor Jason Millman
Department of Education
Stone Hall
Cornell University
Ithaca, NY 14853

1 Bill Nordbrock
Instructional Program Development
Building 90
NET-PDCD
Great Lakes NTC, IL 60088

1 Dr. Melvin R. Novick
356 Lindquist Center for Measurment
University of Iowa
Iowa City, IA 52242

1 Dr. Jesse Orlansky
Institute for Defense Analyses
400 Army Navy Drive
Arlington, VA 22202

1 Dr. James A. Paulson
Portland State University
P.O. Box 751
Portland, OR 97207

1 MR. LUIGI PETRULLO
2431 N. EDGEWOOD STREET
ARLINGTON, VA 22207

1 DR. DIANE M. RAMSEY-KLEE
R-K RESEARCH & SYSTEM DESIGN
3947 RIDGEMONT DRIVE
MALIBU, CA 90265

Non Govt

1 MINRAT M. L. RAUCH
P II 4
BUNDESMINISTERIUM DER VERTEIDIGUNG
POSTFACH 1328
D-53 BONN 1, GERMANY

1 Dr. Mark D. Reckase
Educational Psychology Dept.
University of Missouri-Columbia
4 Hill Hall
Columbia, MO 65211

1 Dr. Andrew M. Rose
American Institutes for Research
1055 Thomas Jefferson St. NW
Washington, DC 20007

1 Dr. Leonard L. Rosenbaum, Chairman
Department of Psychology
Montgomery College
Rockville, MD 20850

1 Dr. Ernst Z. Rothkopf
Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

1 Dr. Lawrence Rudner
403 Elm Avenue
Takoma Park, MD 20012

1 Dr. J. Ryan
Department of Education
University of South Carolina
Columbia, SC 29208

1 PROF. FUMIKO SAMEJIMA
DEPT. OF PSYCHOLOGY
UNIVERSITY OF TENNESSEE
KNOXVILLE, TN 37916

1 DR. ROBERT J. SEIDEL
INSTRUCTIONAL TECHNOLOGY GROUP
HUMRRO
300 N. WASHINGTON ST.
ALEXANDRIA, VA 22314

Non Govt

1 Dr. Kazuo Shigemasu
University of Tohoku
Department of Educational Psychology
Kawauchi, Sendai 980
JAPAN

1 Dr. Edwin Shirkey
Department of Psychology
University of Central Florida
Orlando, FL 32816

1 Dr. Robert Smith
Department of Computer Science
Rutgers University
New Brunswick, NJ 08903

1 Dr. Richard Snow
School of Education
Stanford University
Stanford, CA 94305

1 Dr. Robert Sternberg
Dept. of Psychology
Yale University
Box 11A, Yale Station
New Haven, CT 06520

1 DR. PATRICK SUPPES
INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES
STANFORD UNIVERSITY
STANFORD, CA 94305

1 Dr. Hariharan Swaminathan
Laboratory of Psychometric and Evaluation Research
School of Education
University of Massachusetts
Amherst, MA 01003

1 Dr. Brad Sympson
Psychometric Research Group
Educational Testing Service
Princeton, NJ 08541

Non Govt

1 Dr. Kikumi Tatsuoka
Computer Based Education Research Laboratory
252 Engineering Research Laboratory
University of Illinois
Urbana, IL 61801

1 Dr. David Thissen
Department of Psychology
University of Kansas
Lawrence, KS 66044

1 Dr. Robert Tsutakawa
Department of Statistics
University of Missouri
Columbia, MO 65201

1 Dr. J. Uhlaner
Perceptronics, Inc.
6271 Variel Avenue
Woodland Hills, CA 91364

1 Dr. Howard Wainer
Division of Psychological Studies
Educational Testing Service
Princeton, NJ 08540

1 Dr. Phyllis Weaver
Graduate School of Education
Harvard University
200 Larsen Hall, Appian Way
Cambridge, MA 02138

1 Dr. David J. Weiss
N660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455

1 DR. SUSAN E. WHITELY
PSYCHOLOGY DEPARTMENT
UNIVERSITY OF KANSAS
LAWRENCE, KANSAS 66044

1 Wolfgang Wildgrube
Streitkraefteamt
Box 20 50 03
D-5300 Bonn 2
WEST GERMANY